)RT DOCUMENTATION PAGE

1a.	1b. RESTRICTIVE MARKINGS
AD-A198 214	NA
2a	3. DISTRIBUTION / AVAILABILITY OF REPORT
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE NA	Distribution Unlimited; Approved for Public Release

4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S)
INDU/DC/GMH/TR-88-38	NA

6a. NAME OF PERFORMING ORGANIZATION	6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
Indiana University	NA	ONR

6c. ADDRESS (City, State, and ZIP Code)	7b. ADDRESS (City, State, and ZIP Code)
Department of Chemistry Bloomington, IN 47405	800 N. Quincy Street Arlington, VA 22217

8a. NAME OF FUNDING/SPONSORING ORGANIZATION	8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
		Contract N00014-86-K-0366

8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS			
	PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO. R&T Code 4134006	WORK UNIT ACCESSION NO

11. TITLE (Include Security Classification)

Quantile BEAST Attacks the False-Sample Problem in Near-Infrared Reflectance Analysis

12. PERSONAL AUTHOR(S)
Robert A. Lodder and Gary M. Hieftje

13a. TYPE OF REPORT	13b. TIME COVERED FROM ___ TO ___	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical		11 July 1988	74

17.	COSATI CODES		18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Near-Infrared Reflectance Analysis, Solid-sample analysis, False Samples, Chemometrics, (jhdl)

19. ABSTRACT (Continue on reverse if necessary and identify by block number) Infrared Spectroscopy

The multiple linear regression approach typically used in near-infrared calibration yields equations in which any amount of reflectance at the analytical wavelengths leads to a corresponding composition value. As a result, when the sample contains a component not present in the training set, erroneous composition values can arise without any indication of error. The Quantile BEAST (Bootstrap Error-Adjusted Single-sample Technique) is described here as a method of detecting one or more "false" samples. The BEAST constructs a multidimensional form in space using the reflectance values of each training-set sample at a number of wavelengths. New samples are then projected into this space and a confidence test is executed to determine whether the new sample is part of the training-set form. The method is more robust than other procedures because it relies on few assumptions about the structure of the data; therefore, deviations from assumptions do not affect the results of the confidence test. Keywords:

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	Distribution Unlimited

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code)	22c. OFFICE SYMBOL
Gary M. Hieftje	(812) 335-2189	

88 8

OFFICE OF NAVAL RESEARCH

Contract N14-86-K-0366

R&T Code 4134006

TECHNICAL REPORT NO. 38

QUANTILE *BEAST* ATTACKS THE FALSE-SAMPLE PROBLEM

IN NEAR-INFRARED REFLECTANCE ANALYSIS

by

Robert A. Lodder and Gary M. Hieftje

Indiana University
Department of Chemistry
Bloomington, Indiana 47405

Accession For

NTIS GRA&I	X
DTIC TAB	☐
Unannounced	☐
Justification	

By_____
Distribution/
Availability Codes

| Dist | Avail and/or Special |

A-1

11 July 1988

QUALITY INSPECTED 2

Near-infrared diffuse reflectance spectrometry is a rapid analytical method that typically uses the reflectance of a sample at several wavelengths to determine the sample's composition[1]. The technique is heuristic in its approach and makes extensive use of computers[2,3]. Through a computational modeling process (generally employing multiple linear regression), near-infrared reflectance analysis is able to correct automatically for background and sample-matrix interferences, making ordinarily difficult analyses seem routine. The modeling process employs a "training set" of samples to "teach" the computer algorithm to recognize relationships between minute spectral features and the sample's composition[4]. Of course, the training set must have been previously analyzed by some other reliable (reference) chemical procedure. Although assembling a training set and developing a new calibration can require considerable time, the subsequent speed of quantitative analysis has provided plenty of impetus for the growth of near-IR reflectance methods.

Quantitative analysis has been the principal application of near-IR reflectance analysis to date[5]. Recently, however, some attention has been turned to the use of near-IR reflectance as a qualitative technique as well[6-10]. Near-IR reflectance analysis has been shown to be capable of differentiating among a variety of pure compounds and mixtures of constant composition. It is this ability that is exploited here to solve the false-sample detection problem.

A false sample is simply any sample that falls outside of the domain of the samples used to train the near-IR reflectance analysis algorithm. For example, a manufacturer may be interested in using near-IR reflectance analysis to monitor the protein concentration of a liquid

stream. The normal range of concentrations might be 3 to 6%, and training samples would be selected to completely cover this range. If a process change or equipment failure should occur one day and the protein concentration jump to 10%, a false-sample situation would exist. Analyzing this false sample requires extrapolating beyond the range of the training set used to generate the prediction equation. An operator should be signaled either to stop the stream and correct the equipment failure, or to recalibrate the near-IR reflectance analysis instrument to accept the new range of concentration values.

This type of false-sample condition is easily detected, however, by a simple test to determine if the predicted value falls outside of the range of concentrations used in generating the prediction equation. Another type of false-sample condition is more insidious and difficult to detect. A completely new component, a component not present in the training set and therefore thoroughly unexpected, can appear in the samples and cause erroneous composition values to be generated. This component could be a chemical entity, as might be introduced by opening a valve at the wrong time or by contamination of the raw materials, or from a noise source, such as instrument drift over time or a change in particle-size distribution. In short, the aim of false-sample detection is to go beyond simple qualitative analysis to answer the question, "Does my prediction equation apply to the current sample?"

The process of detecting false samples involves the analysis of multivariate data distributions, a topic which is currently being investigated in a number of ways[11]. We have selected quantile analysis[12] as a basis for nonparametric tests of distributional assumptions because it provides easy access to both numerical statistics and readily

-4-

interpreted graphs. Quantile analysis simply transforms the cumulative frequency distribution of a data set into a convenient linear form. From this form the location, scale, and skew of the data can be estimated. Quantile analysis provides additional advantages[13] that are particularly useful with multivariate data. These advantages include the following:

1. The complexity of the graph is independent of the number of observations.

2. The quantiles are invariant under monotone transformation (such as a transformation of a distribution location or scale).

3. Condensation, interpolation, and smoothing of data are easily accomplished.

4. The grouping difficulties that occur in histograms are not present.

5. Peculiarities, such as overlap of two distributions or multimodality, are effectively indicated.

Figure 1 shows reflectance data obtained from a number of spectra, using two wavelengths to describe two hypothetical compounds, A and B. (Each wavelength in a spectrum can be represented as a spatial dimension, giving a single point in an n-dimensional space for a spectrum recorded at n wavelengths. The point is translated from the origin by amounts that correspond to the magnitude of the reflectance observed at each wavelength. By representing spectra in this manner, a group of similar samples with similar spectra appears as a cluster of points at a certain location in space.) A univariate distribution can be formed from the points that lie within a specified radius of the line connecting the centers of clusters A and B.

Quantile plots are often utilized in the analysis of univariate distributions to compare a theoretical distribution to an empirical one. A particular quantile (p) selected for plotting represents the value of the integral of a probability density function (from negative infinity until the quantile p is reached). Comparing two distributions by their quantiles requires that both distributions be transformed into cumulative distributions. Essentially, this is accomplished by integrating the two probability density functions to form a theoretical cumulative distribution function (TCDF) and an empirical cumulative distribution function (ECDF). Typically, the upper limits of integration for p are what is plotted for the TCDF and the ECDF. Starting with p and solving for a limit of integration constitutes the inversion of the distribution function. By convention, quantile plots put the TCDF on the x-axis and the ECDF on the y-axis. The scales used on each axis are derived from the values of the corresponding inverse cumulative distribution functions as p is allowed to vary between zero and one[13].

Figure 2 shows a quantile plot of the points along the center line (inside the box) from Figure 1. The inset in Figure 2 is a histogram (empirical distribution function) of the same points. These points, used to form the ECDF and set along the ordinate, are plotted versus the quantiles of the normal distribution (the TCDF) on the abscissa. The slopes and intercepts of the two lines in Figure 2 supply parameters to equations for the probability density in the direction through the two cluster centers. In a similar manner, quantiles can be used to set confidence limits around clusters by determining the probability density in a specified direction.

Multivariate data analysis using distribution quantiles provides a way both qualitatively to identify samples and to determine when a new near-IR reflectance calibration equation is required. The algorithm is conceptually uncomplicated and takes a step toward simplifying the statistics of near-IR reflectance analysis in the manner in which near-IR reflectance methods themselves have simplified instrumentation. A description of this algorithm is the subject of the following section.

THEORY

Data Clusters in Near-IR Reflectance Analysis. The sample-identification problem in near-IR reflectance analysis can be a complex one indeed, and time can be spent profitably in the examination of a fairly simple construction of this problem. Figure 3 depicts the spectra of three hypothetical pure compounds, A, B, and C. These spectra were recorded at two wavelengths and projected into a two-dimensional space as described in the previous section. The measurements are assumed to be free of error from particle-size differences, concentration variations, drift, etc., and therefore result in three points in space rather than three clusters. When 1000 sample mixtures are prepared from A, B, and C by randomly weighting the proportion of each compound in each mixture, a training set (whose spectra are shown in Figure 4) is formed. (Figure 4 assumes that Beer's Law holds.) This training-set cluster is basically elliptical; should one desire to determine the distance of a new sample from this cluster, the Mahalanobis metric[14] provides an obvious means. Figure 5 shows the resulting training set when compounds A and C are moved approximately three times farther away from B than they were in

Figure 4. The training set remains elliptical, but its orientation has been altered (the slope of the line through the major axis of the ellipse has increased, as shown by the scales on the x and y axes). This behavior demonstrates that the shape of the training-set cluster is basically independent of the positions (spectra) of the raw materials (A, B, and C), at least in the error-free case represented by the pure components depicted in Figure 3.

Figure 6 corresponds to Figure 3, except that the pure-component spectra are no longer precisely known. A, B, and C are now 1000-point clusters, formed by adding bivariate-normal noise to the original three points. Each of these clusters is also elliptical. The variance of the major (horizontal) axis is arbitrarily set four times greater than that of the minor (vertical) axis. When a randomly weighted training set of 1000 hypothetical mixtures is created from the clusters corresponding to the points used in Figure 5, a revealing pattern (Figure 7) emerges. The smooth elliptical shape of the cluster has broken down and the distribution of points about the center of the cluster is no longer perfectly symmetric. A cross-section of points through the center of the cluster shows a definite right(or positive)-skew, indicating the presence of a leverage effect by points with higher values. (A frequency-distribution polygon of this cross-section appears in Figure 8.) Real near-IR reflectance data showing this same effect appears in Figure 18 and will be discussed later.

Figure 9 demonstrates the effect of such a skew on the determination of a confidence limit. Because of the balancing property of the mean (μ), a small number of points located some distance away from the others exerts a considerable leverage on the value of the mean. When these

-8-

distant points are lower values than the rest as in Figure 9, the distribution is said to be left(or negatively)-skewed, and the mean shifts downward. In Figure 9, these outlying points are real and should not be discarded (the solid line represents the underlying distribution of the total population from which the samples were drawn). For this situation, the usual symmetric statement of confidence limits[15], e.g.,

$$\text{(confidence limits for } anything\text{)} = \mu +/- t_{\alpha,\phi} s_{anything}$$

does not provide an adequate description. For example, suppose $z_\alpha \sigma = a$ (see Figure 9). Clearly, the probability of a point appearing at $(\mu+a)$ is different than the probability of a point appearing at $(\mu-a)$.

Discriminant Analysis in Near-IR Reflectance. Asymmetry of near-IR spectral clusters about their means has appeared in literature reports[9]. However, the methods that have been applied to discriminant analysis in near-IR reflectance have not dealt with this phenomenon. In general, discriminant analysis involves several assumptions[16], including:

1. No discriminating variable is a linear combination of other discriminating variables.

2. The covariance matrices for all (spectral) groups are approximately equal (unless special formulas are used).

3. Each group has been drawn from a population that is normally distributed on the discriminating variables.

The violation of these assumptions reduces the efficiency of discriminant analysis, and increases the probability of misclassifying samples.

The Quantile BEAST. Nonparametric methods (distribution-free methods
whose properties hold under very few assumptions about the population
from which the data are obtained) can be profitably applied to this sort
of discriminant analysis.   Nonparametric techniques   provide   a
conceptually simpler statistical alternative to many common procedures,
and for the price of some additional arithmetic, can keep violations of
assumptions from being reflected in faulty inferences.   The Quantile
BEAST (Bootstrap Error-Adjusted Single-sample Technique) is such a
distribution-free method of flagging samples outside of the domain of the
training set. It is based on the bootstrap procedure of Efron[17], a
nonparametric method of assigning a standard error to a point estimate.
The Quantile BEAST constructs a multidimensional cluster in space using
the reflectances of each training-set sample at a number of wavelengths.
New samples are then projected into this space and a nonparametric
confidence test is performed to determine whether the new sample is part
of the training-set cluster.   In this manner, qualitative identification
of pure samples is made possible and false-sample mixtures can be
detected.

Whenever it is possible that the distribution underlying a set of
samples is skewed, an   asymmetric nonparametric method of setting
confidence limits should be employed[19].   This sort of method does not
assume that the best point estimate of the mean lies at the center of the
interval between the confidence limits.   The BEAST acquires this
asymmetric estimation capability from the bootstrap, which can be
summarized as follows. Given an unknown distribution function F, and
some parameter of interest (such as the sample mean or median) that is a
function of a number of independent identically distributed observations

from F, the object of the bootstrap is to determine the standard error in the parameter of interest from the observations in a training set (a representative sampling of the unknown distribution). The real standard error of the parameter is a function of the unknown distribution, the size of the sample set, and the form of the parameter. However, knowing the number of observations and the form of the parameter, the standard error can be expressed as a function of only the unknown distribution F (in essence, this is the definition of the bootstrap estimate of the standard error). In other words, although the actual distribution F is unknown, we can estimate it using the empirical probability distribution represented by the training set. There are three steps to this estimation process:

1. A training set composed of equally weighted observations from the unknown distribution is created. This empirical probability distribution (see Figure 10b) must be constructed to adequately describe the variation in the unknown distribution F (Figure 10a). The 19 observations that appear in Figure 10b can be thought of as 19 "blocks" removed from an infinitely large pile of similar observation "blocks" (Figure 10a). These training-set blocks are stacked in Figure 10b in a manner that describes their original positions in the infinitely large pile, thus forming a histogram. Once assembled, this training set remains unchanged for the remainder of the estimation procedure.

2. Bootstrap observations are blocks that are randomly and independently drawn from the training set, with replacement from the training set (so that the training set is never depleted or changed), to form a bootstrap set with the same total number of blocks as the training

-11-

set (19 blocks for each set in Figures 11a-c). Note that the number of observations appearing at a given location in a bootstrap set will vary between bootstrap sets taken from the same training set.

3. The sampling distribution of the parameter of interest (e.g., the center of a group of spectral points in hyperspace) for the unknown distribution F is approximated by the distribution of that parameter calculated using the bootstrap observations (hereafter this distribution is termed simply the "bootstrap distribution").

The most difficult part of the bootstrap procedure is the actual calculation of the bootstrap distribution. There are three ways to arrive at this distribution.

1. Direct theoretical calculation of the distribution can be performed. This works for a few special examples, but is usually impossible in experimental situations.

2. Taylor-series expansion methods can be used to find an estimate of the mean and variance of the bootstrap distribution of the parameter of interest. This turns out to be a form of the jackknife, another nonparametric method[17].

3. The Monte-Carlo approximation to the bootstrap distribution can be calculated. With this method, a large number (B) of bootstrap sample sets is generated by randomly drawing observations from the training set (again, with replacement from the training set) to form bootstrap sets of the same size as the training set. The empirical distribution of the corresponding values of the parameter of interest calculated using the bootstrap sets is taken as an approximation to the bootstrap distribution. (This process is shown in Figures 11a-c, where the three figures represent B=3 bootstrap replications of the

-12-

training set, and in Figure 12, which shows a Monte Carlo approximation to the bootstrap distribution.) The Monte Carlo method is by far the most commonly used experimental technique for obtaining the bootstrap distribution.

The bootstrap approach illustrates a way in which data analysis in near-IR reflectance spectroscopy can be simplified, much like near-IR reflectance methods themselves have simplified spectroscopic measurements. Specifically, a complex problem can be solved by a simple procedure iterated a large number of times. In effect, each sample spectrum in the training set is simply copied thousands of times. The resulting spectra are then thoroughly shuffled and new training sets are selected from the copies at random. An expectation value (the center of the spectral-point cluster in hyperspace) is calculated from each new training set. The distribution of bootstrap-set centers can be treated as though it were constructed from actual collected training sets, and its quantiles can be used to produce an estimate of the value and precision of the center-parameter for the original population. The quantiles are therefore useful in defining the boundaries of a training set or pure compound in the hyperspace of spectral points.

Bootstrap distribution quantiles are readily converted into confidence intervals. Using the method described above, in which the distribution of bootstrap-set centers is treated as though it were constructed from actual observations, selecting any two bootstrap-distribution percentiles gives the corresponding confidence limits for the center-parameter (e.g., selecting the 16th and 84th percentiles of the bootstrap distribution produces the central 68% confidence limits). At times, however, this method can fail to capture the asymmetry of a

-13-

distribution. (The failure becomes obvious in near-IR spectral data when the median of the training set is noticeably different from the mean of the bootstrap distribution.) Similar observations led Efron to propose a bias correction method that was largely incorporated into the BEAST (forming the Error-Adjustment)[18]. In essence, applying the correction is a recognition that the mean value of a distribution drifts in the direction of the skew (measured with respect to the median of the distribution). Large differences between the median and mean suggest a skew that casts doubt upon the validity of a simple symmetric confidence interval. When no difference between the mean and the median exists, the value of the correction is zero. Otherwise, the error is compensated and new confidence limits are obtained with an improved representation of any skew present.

The entire BEAST algorithm can now be outlined. In the technique, each monitored wavelength is considered a dimension in hyperspace and the distribution of reflectances on each wavelength axis gives projections of the clusters of points. Each point represents an entire spectrum, translated from the origin by amounts that correspond to the magnitude of the reflectance observed at each wavelength. Valid samples are defined as those that fall inside the cluster of training-set points, while false samples are those that fall outside of the cluster. Confidence limits, set along any linear combination of wavelengths (dimensions), define the surface of the cluster at a specified confidence level. These confidence limits are obtained by using the bootstrap (hence the B in the BEAST) to arrive at an estimate of the real-sample population distribution based upon the training-set distribution. The center of the real-sample distribution is estimated by the BEAST using the center of the bootstrap-

set centers (the bootstrap distribution). When a new sample (the Single-sample in the BEAST) is tested, a vector is formed in hyperspace between the new spectral point and the estimated center point of the real-sample distribution. A hypercylinder formed about this line will contain a number of estimated real-sample training-set centers. When the coordinates of these points are transformed into distances from the estimated center of the real-sample distribution, a univariate distribution is constructed. It is this univariate distribution that is used in the confidence test. The reliance on nonparametric techniques produces a false-sample test that operates without assumptions about the shape, size, symmetry, or orientation of the spectral-point cluster in hyperspace.

EXPERIMENTAL

The algorithms previously described were implemented in programs using VAX-11 BASIC (version 2.4, Digital Equipment Corporation) and Speakeasy IV Delta (VMS version, Speakeasy Computing Corporation, Chicago). These programs were run on VAX-11/780 and VAX-11/785 computers. Spectral data at 18 discrete wavelengths were collected using a Technicon InfraAlyzer 400 filter spectrophotometer. This spectrophotometer was directly connected to a VAX-11/780 computer using custom interface and graphics programs.

Three types of demonstrations were conducted using the BEAST, beginning with the common compound-identification problem and proceeding to the more difficult problem of detecting contaminated mixtures. The three tests were:

-15-

1. The BEAST was used qualitatively to distinguish among a number of pure benzoic-acid derivatives. Each of the four compound names in Table I represents three spectra of that compound obtained by rotating the closed-sample cup in the drawer and scanning through the 18 filters.

2. The BEAST was trained with a set of 40 mixtures of the benzoic acid derivatives used in test 1. The samples were examined at three selected wavelengths, and the BEAST was then presented with pure benzoic acid derivatives and other compounds to determine if it could detect these false samples. A total of 60 samples were prepared for the training set in this test; of these, 20 were retained for use in test 3 below. More specifically, the second type of BEAST demonstration involved developing a training set composed of random mixtures of salicylic, benzoic, isophthalic, and p-aminobenzoic acids. Each of these four components was allowed to vary from 0 to 25% (by weight) in each sample. The remainder was made up of the aluminum-oxide diluent. A random-mixing algorithm (the same one used to generate the theoretical training set of Figures 4, 5, and 7) was used to generate the amounts to be used for each component in each sample. The purpose of this test was to develop a large training-set spectral cluster that contained variations from a number of sources, and then to determine whether pure samples could still be correctly identified as being different (or false) using only the training-set spectral data. The results for a few compounds from the laboratory shelf appear in Table II.

3. The 40-sample training set from test 2 was used again, this time in conjunction with the 20 mixture samples that were held out of the

-16-

training set created in test 2. Ten of these 20 samples were "contaminated" with a false-sample compound to determine whether these samples would be flagged. The other ten were unaltered and served as validation samples.

The benzoic-acid derivatives used in test 1 and in the training set for tests 2 and 3 were analytical reagent-grade salicylic acid, p-aminobenzoic acid, isophthalic acid, and benzoic acid. Reagent-grade aluminum oxide was used as a diluent in mixtures where a range of component concentrations was desired (tests 2 and 3). Before each sample (or sample mixture) was read in the InfraAlyzer 400 spectrophotometer, the sample was ground and mixed in a Spex mixer/mill. The powder was then sifted through a 100-mesh sieve and packed into the closed sample cup provided with the InfraAlyzer. Three readings were taken on each sample, each successive reading after a 120-degree rotation of the closed cup.

RESULTS AND DISCUSSION

Results of Studies Using Benzoic-Acid Derivatives as Samples. The ability of the BEAST to differentiate among four rather similar benzoic-acid derivatives (test 1) is suggested by the data in Table I. The distances tabulated have units of asymmetric nonparametric central 68% confidence intervals (equivalent to one standard deviation if the underlying distribution were Gaussian). These intervals were calculated from the distance (Euclidean metric) between the 0.16 and 0.84 quantiles of the bootstrap distribution, following the projection of this distribution onto the hyperline connecting the center of the bootstrap

-17-

distribution and the new-sample spectral point. These compound-spectra form four clusters in hyperspace whose intercluster distances (expressed in "standard deviations") appear in Table I. The compound heading each column was designated as the training set, so that the distance from a column-heading compound to a row-heading compound is given in terms of nonparametric standard deviations (SDs) of the column-heading compound. For example, the distance between the clusters representing benzoic acid and salicylic acid can be expressed in two ways: in terms of the standard deviation of benzoic acid in the direction of salicylic acid (39 SDs), or in terms of the standard deviation of salicylic acid in the direction of benzoic acid (137 SDs). The difference in SDs, of course, reflects the difference in the variances (sizes) of the benzoic and salicylic acid clusters in the direction of each other.

In fact, benzoic acid exhibited the largest overall variance of any of the compounds tested, because of the difficulty of grinding benzoic acid, which forms long, thin crystals. Such crystals can slip through a 100-mesh sieve and thereby introduce an additional source of variation into the benzoic acid data cluster. The variation in the other compounds can be attributed primarily to orientation effects caused by packing peculiarities, because this was the only factor that was varied between replicate measurements of these substances. Aluminum oxide (included in this test because it was used as the diluent in tests 2 and 3) had the smallest overall variance, not surprising since it was the only reagent that was available in a powder fine enough to pass through the 100-mesh sieve.

Other distances in Table I are also in accordance with what might be predicted: aluminum oxide and isophthalic acid are 4208 SDs apart,

because chemically they have little in common. In fact, aluminum oxide is the most distinct from all of the other compounds. Benzoic acid and salicylic acid form the closest spectral clusters, since they differ by only a single oxygen atom. For all these pure components, only a small number of bootstrap replications are required (B=50) because the variance is relatively small and the distance between adjacent clusters (measured in 18-dimensional space) is relatively large. This clear disparity allows the samples to be distinguished computationally in a fraction of a second. Considering the distances (in terms of SDs) involved, there is little danger of any of these pure compounds being incorrectly identified by the BEAST.

Benzoic-Acid Derivative Mixtures. A large number of sample components, varying over a relatively broad range of concentrations, can have the effect of "filling" the spectral hyperspace provided by only a few analytical wavelengths. This phenomenon is demonstrated by the 3-wavelength training set used in obtaining Table II. The Euclidean distances for the data in Tables I and II were quite similar (all within an order of magnitude or so of each other); however, the distance in terms of standard-deviation units is far more disparate between the two tables. The reason that the distances in SDs shrink so much from Table I to Table II is that the cluster size in Table II has increased by about a factor of 100. This spread causes some apparently unrelated compounds, like sucrose and whole wheat flour, to appear to be similar to the training set of benzoic acid derivatives. This increase in training-set cluster size also highlights the need to to be aware of the exact shape of the training-set cluster when small distances are classified as representing valid or false samples. Finally, mixing a number of

-19-

different components results in diluting the contribution of each one to the total sample spectrum, in effect compressing the available analytical space further and increasing the probability of mixture-clusters overlapping. (This behavior can be seen in Figures 3 and 4 as well as 6 and 7.)

Contaminated Bensoic-Acid Derivative Mixtures. The third test of the BEAST investigated this worst-case mixture-overlap situation by employing the same training set described in test 2 above to train the BEAST algorithm. The algorithm was then presented with the remaining 20 sample mixtures. Ten of these 20 mixtures were "contaminated" with acetylsalicylic acid (randomly varying over a range from 1 to 20%). The results appear in Table III. Even in this worst-case example it is apparent that the contaminated samples are likely to be detected as false. Three out of ten of these contaminated samples fail the 3 SD test for being false. None of the uncontaminated samples was incorrectly identified.

One might wonder why the response (in terms of SDs) for very similar contaminant concentrations varies so widely among contaminants in Table III and why contaminant concentration does not appear to correlate well to distance. The answer seems to be in what the other four component concentrations are. When components similar to acetylsalicylic acid rise in concentration, a 1% contribution to the final spectrum from acetylsalicylic acid becomes relatively smaller and the sample spectral point appears to move closer to the training set. Table IV shows the distance response of the BEAST for two groups of similar acetylsalicylic-acid contaminant concentrations. The concentration of the diluent (aluminum oxide) is inversely related to the concentrations of the

-20-

benzoic acid derivatives, so lower diluent concentrations indicate an increase in the concentrations of the other (non-contaminant) compounds. This relative increase in the concentrations of the noncontaminant compounds, in turn, correlates to a decrease in the distance of the sample from the training set.

Results of Theoretical Studies Using Synthetic Samples. A number of parameters affect the performance of the BEAST in experimental situations, including the number of wavelengths used in the analysis, the training-set size, the selected radius of the hypercylinder, and the number of bootstrap replications of the training set that are employed. To investigate the effects of each of these variables, a theoretical calculation was undertaken using hypothetical multivariate samples, randomly drawn from a multivariate normal population with a known group mean (center) and a known variance in all directions. A number of these synthetic samples (selected by Monte-Carlo integration of the multivariate normal population) formed a training set that was then analyzed by the BEAST, whose task it was to indicate the variance of the training set in a selected direction. The bias and mean-square error (MSE) of the BEAST as a point estimator of the variance could then be determined for a particular combination of parameters. Ten runs were made with each combination of parameters (1, 2, 3 and 5 wavelengths, training set sizes from 10 to 200 samples, hypercylinder radii from 0.001 to 0.1, and 50, 200, 1000, and 10000 bootstrap replications). The results are summarized in a series of figures that describe the bias and error of the BEAST estimator as a function of each of the four parameters. The figures serve as a brief guide to using the BEAST by providing estimates of the error that can be expected for typical

-21-

combinations of training-set size, number of replications, number of wavelengths, and hypercylinder radius.

The most influential factor affecting the bias, or accuracy, of the BEAST appears to be the size of the training set. Figure 13 depicts the percentage bias (given by the absolute value of $100(E(t)-K)/K$, where K is the known value used to generate the hypothetical samples and $E(t)$ is the estimate of the true value given by the BEAST) of the BEAST as a function of the training-set size for a typical run (d=2 wavelengths, B=1000 bootstrap replications). The bias is quite large (28.5%) when only 10 training samples are used, but it drops steadily to under 1% when the training set size reaches 200 samples. This suggests that to be prudent one must use as many training samples as possible, a result that is not startling in light of the requirements of other near-IR methods[20]. A particularly poor training set (e.g., a small number of samples tightly clustered in a small region in spectral hyperspace) can cause the bias to jump to over 106% (as occurred in a run where d=2 wavelengths, B=10000 replications, and n=10 training samples). Potential users of the BEAST are therefore cautioned to make certain that their training samples are well-distributed, and not assume that they are well-distributed simply because they were randomly selected. Histograms[13], sample-selection algorithms[21], and quantile-quantile plots[13] all provide ways for users to ensure the adequacy of a particular training-set distribution

Figure 13 also shows the relationship between the mean-square error (MSE) and the training-set size. The MSE has been converted to relative standard deviation (RSD) and expressed as a percentage. Overall the RSD clearly decreases (improves) as the training-set size increases (the slight increase in RSD at n=200 samples is probably an artifact of the

-22-

particular sample set that was randomly selected). The training-set size does not influence the RSD to the same degree as does the number of bootstrap replications of the training set (shown in Figure 14).

The number of bootstrap replications of the training set is the most influential factor in determining the RSD of the BEAST. Figure 14 gives the RSD as a function of the number of bootstrap replications for sampled spectra consisting of 1, 3, and 5 wavelengths. Two things should be noted from these plots: first, that the RSD drops rapidly as more bootstrap replications are performed, and second, that using more wavelengths in the analysis demands the use of more replications to achieve a given RSD. This behavior is not surprising. The BEAST uses the bootstrap-replicate distribution to approximate the real-sample distribution. Therefore, more replicates are required as more wavelengths are used because the size of the analytical space that must be described by these replicates increases. The number of replicates must be large enough to assure an adequate number of points in the hypercylinder as well (though as few as 50 points in the hypercylinder are often sufficient). Of course, to a certain extent, the number of points in the hypercylinder can be increased by increasing the radius of the hypercylinder. Eventually, however, this approach results in a loss of directional selectivity that begins to bias the quantiles of the data in the hypercylinder. Figure 15 depicts the percentage bias of the BEAST as a function of the number of bootstrap replications of the training set. The accuracy is not as strongly affected as is the RSD by a small number of replications.

The effect of the hypercylinder radius on the RSD and bias is shown in Figure 16. The relatively high RSD values in the figure are the

-23-

result of using only 50 bootstrap replications to describe the synthetic data in a five-dimensional space. As mentioned above, the number of points in the hypercylinder can be controlled by changing the radius of the hypercylinder. Many rules for setting this radius could be proposed, such as setting the radius by a function of the average nearest-neighbor distance, or by a function of the distance of the smallest dimension in the cluster. In practice, the best results were most easily obtained by keeping the radius of the hypercylinder two to three orders of magnitude smaller than its length. In our experience this approach provides an adequate number of points inside the hypercylinder and preserves sufficient directional selectivity.

Figure 17 shows the effect of the number of wavelengths on the RSD and bias. When the hypercylinder radius is controlled (as was done in this figure) the bias and RSD will vary only slightly and will remain at low levels. One would expect both lines to curve upward at a larger number of wavelengths (if additional replicates are not used). Of course, if the number of replications and the hypercylinder radius are not controlled, fewer and fewer points fall inside the hypercylinder as the number of wavelengths increases. In this event, the results of the BEAST begin to deteriorate rapidly. The effect of the hypercylinder radius is particularly sensitive to the number of wavelengths that is used (the radius had to be increased from 0.001 for the three-wavelength test runs to 0.005 for the five-wavelength test runs to get any points inside the hypercylinder at all).

-24-

CONCLUSIONS

Conventional near-infrared reflectance analysis correlates changes in sample spectra to changes in sample composition. In this sense, the near-IR analysis algorithm is essentially a pattern-recognition technique that produces a linear equation relating a particular component concentration to reflectances observed at several near-infrared wavelengths. Like most pattern-recognition methods, the near-IR reflectance analysis algorithm is accurate only as long as the samples presented to it are what it "expects to see." When a new sample falls outside of the domain of the sample set used to train the near-IR reflectance analysis algorithm, the near-IR prediction equation does not apply to the new sample. The Quantile BEAST provides a way to adequately detect this false-sample situation. The new sample can then be rejected or the algorithm can be retrained with samples known to be similar to the new (false) one.

In general, the shape and orientation of experimentally obtained multidimensional near-IR spectral data are not completely predictable[7]. This fact is suggested by both the synthetic data in Figure 7 and by the experimental data in Figure 18. Although this unpredictability is usually not a problem when the simple identification of pure compounds (or mixtures of low variability) is attempted, it can become a problem when:

1. the allowable variability of the mixtures increases (e.g., from drift over time, changing component concentrations, or variations in particle size or sample packing), or

-25-

2. mixture samples are contaminated by a foreign component not present in the training set.

A nonparametric clustering method, the Quantile BEAST, functions without assumptions about the shape, size, orientation, or symmetry of the data. Deviations from assumptions cannot be reflected in the method's results. Like many other nonparametric methods, the BEAST is based on a very simple procedure (the drawing of a bootstrap sample) iterated a large number of times to produce a whole result (the sample classification) that is seems almost greater than the sum of the method's parts. One of the strengths of the BEAST is its ability to produce an understandable result in an understandable way.

The false-sample problem is an important one with many facets, and lives can literally depend on getting a correct answer. False samples might be over-the-counter drug capsules filled with poison, and placed in a batch of the unadulterated drug capsules. False samples might be decoy missile warheads dispersed among real warheads, traveling in space en route to their targets. The BEAST is potentially applicable to both situations, and to more. However, the BEAST is not limited merely to use in solving the false-sample problem. It is easy to envision using the BEAST to select training-set samples as well. In many near-IR reflectance analysis applications it is desirable to include a relatively large number of "extreme" samples in the training set in order to obtain a prediction equation with a uniform error over its entire range. The BEAST can take a large number of potential training samples in wavelength space and create an estimate of the true training set using bootstrap-probability space. A uniform sampling of the potential training samples in wavelength space (using each sample's probability of belonging to the

training set, obtained from the bootstrap space) will then produce a representative training set from the spectra without the need for tedious chemical analysis of each potential training-set sample.

A weakness of the BEAST that does not hinder parametric methods is the requirement that a bootstrap-replicate distribution be calculated. The largest distribution used in this research (10000 replications) took 2 minutes and 21 seconds of CPU time to create. If this distribution had to be computed using a single CPU for each sample to be analyzed, the BEAST would be impractical in many applications. In practice, however, just as a training set is assembled only once, its replicate distribution need only be created once as well. The important requirement then becomes one of computer memory space and not of time. For many analyses the microcomputers currently available with 256K memory may be adequate.

The BEAST algorithm is naturally suited to implementation in a parallel-processing environment because its bootstrap samples are drawn randomly, independently, and with replacement from the near-IR spectral training set. Basically, the BEAST reaches its solution n times faster when n processors are employed. This parallel-processing mode was used to produce the bias and RSD data for the BEAST that appears in Figures 13-17. The deceptively simple plots disguise the fact that five VAX 11/780 computers were used in parallel to create and analyze over 80000 multicomponent synthetic samples with various combinations of training-set size, hypercylinder radius, and number of wavelengths and bootstrap replications used. After the algorithm was recoded into parallel form the bias and RSD results were obtained in a single afternoon. As parallel processing becomes more common and hardware becomes readily

-27-

available, the BEAST will become an even more attractive option in an ever-increasing number of real-time analytical problems.

CREDIT

-28-

LITERATURE CITED

1. D. L. Wetzel, Anal. Chem. 55, 1165A (1983).

2. D. E. Honigs, T. Hirschfeld, and G. M. Hieftje, Appl. Spectrosc. 39, 1062 (1985).

3. D. E. Honigs, T. B. Hirschfeld, and G. M. Hieftje, Anal. Chem. 57, 443 (1985).

4. P. Rotolo, Cereal Foods World 24, 94 (1979).

5. C. A. Watson, Anal. Chem. 49, 836A (1977).

6 H. L. Mark and D. Tunnel, Anal. Chem. 57, 1449 (1985).

7. H. L. Mark, Anal. Chem. 58, 379 (1986).

8. J. S. Shenk, I. Landa, M. R. Hoover, and M. O. Westerhaus, Crop Sci. 21, 355 (1981).

9. E. W. Ciurczak and T. A. Maldacker, Spectroscopy 1(1), 36 (1986).

10. A. M. C. Davies and W. F. McClure, Analytical Proceedings 22, 321 (1985).

11. P. C. Jurs, Science 232, 1219 (1986).

12. E. Parzen, *Some Recent Advances in Statistics* (Academic Press, London, 1982), pp. 23-52.

13. M. B. Wilk and R. Gnanadesikan, Biometrika 55, 1 (1968).

14. J. A. Hartigan, *Clustering Algorithms* (Wiley, New York, 1975), p. 68.

15. D. G. Peters, J. M. Hayes, and G. M. Hieftje, *Chemical Separations and Measurements* (W. B. Saunders, Philedelphia, 1974), p. 25.

16. W. R. Klecka, *Discriminant Analysis* (Sage Publications, Beverly Hills, 1980), pp. 8-11.

17. B. Efron, Biometrika 68, 589 (1981).

18. B. Efron, Canadian J. Stat. 9(2), 139 (1981).

19. B. Efron, Ann. Statist. 7(1), 1 (1979).

20. D. E. Honigs, G. M. Hieftje, and T. Hirschfeld, Appl. Spectrosc. **38**, 844 (1984).

21. D. E. Honigs, G. M. Hieftje, H. L. Mark, and T. B. Hirschfeld, Anal. Chem. **57**, 2299 (1985).

-29-

TABLE I. Using the BEAST for Qualitative Analysis

      Differentiating Among Four Benzoic-Acid Derivatives

      and a Diluent (distances in BEAST SDs[a])

	Isoh2p[b]	Benzo[c]	Al_2O_3[d]	Salcyl[e]	PABA[f]
Isoh2p	0	188	4208	638	1072
Benzo	1254	0	2889	137	483
Al_2O_3	3930	405	0	1488	3041
Salcyl	1197	39	2991	0	449
PABA	1052	72	3198	234	0

[a]The distance between any two compound-cluster combinations is given in terms of the column-heading compound in the direction of the row-heading compound.

[b]isophthalic acid

[c]benzoic acid

[d]aluminum oxide (the diluent in later tables)

[e]salicylic acid

[f]p-aminobenzoic acid

-30-

TABLE II. Distances (in BEAST SDs) of Four Pure Ground (100 mesh)

Compounds from a Training Set* of Mixtures of Compounds

Compound	Distance
acetylsalicylic acid	8.20
dextrose	46.33
whole wheat flour	12.71
sucrose	3.88

*Training set composed of mixtures of benzoic acid, isophthalic acid, salicylic acid, p-aminobenzoic acid, and an aluminum oxide diluent.

-31-

TABLE III. Distances (in BEAST SDs) of Real and False (contaminated)

Mixtures from the Mixture Training Set Used in Table II

Sample No.	Real Samples[a] (Validation Set)	False Samples[b] (Contaminated Set)	Contaminant (%)
1.	1.29	6.59	1.1
2.	.61	3.25	8.0
3.	.76	5.05	3.8
4.	.47	4.55	14.8
5.	1.97	2.66	4.2
6.	1.54	1.77	19.9
7.	1.65	6.53	1.7
8.	1.18	3.34	1.0
9.	.79	3.88	10.7
10.	1.10	2.32	4.5

(Note: the cluster surface is customarily defined as being three standard deviations from the center of the cluster.)

[a] "Real" mixtures contain benzoic acid, isophthalic acid, salicylic acid, p-aminobenzoic acid, and aluminum oxide. In essence, the real samples form a validation set because they contain the same compounds as the training set samples.

[b] "False" mixtures contain the same components as real mixtures, except that acetylsalicylic acid is also added.

TABLE IV. Distance Response of the BRAST (in standard deviations) for
Two Groups of Similar Contaminant Concentrations in the False
Samples from Table III

Sample No.	Distance(SDs)	Diluent(%)	Acetylsalicylic Acid (%)
1.	6.59	67	1.1
7.	6.53	64	1.7
8.	3.34	50	1.0
3.	5.05	57	3.8
5.	2.66	41	4.2
10.	2.32	44	4.5

Figure Captions

Figure 1. Two thousand spectra, taken at two wavelengths, of two hypothetical compounds, A and B. These spectra are represented as points in a two-dimensional space. A line can be formed between the centers of these two clusters. The box in this figure marks off a region that includes all points within a certain distance (radius) of this center line.

Figure 2. A quantile-quantile (QQ) plot of the spectral data points within the radius of the center line shown in Figure 1. The best-fit straight lines through the two groups of points are shown. The slopes of these lines are the same because the two clusters (A and B) have the same variance in the direction of one another. The inset is a histogram (empirical distribution function) of the points inside the "cylinder" formed about the center line in Figure 1, i.e., the points that generated the QQ plot. The horizontal axis of the histogram covers the same range as the full horizontal axis of the QQ plot.

Figure 3. Two-wavelength spectra of three pure hypothetical compounds, A, B, and C. The spectra are free of error from all sources.

Figure 4. Hypothetical spectra of 1000 training samples, composed of mixtures with randomly selected proportions of A, B, and C, whose two-wavelength spectra are shown in Figure 3.

Figure 5. Hypothetical spectra of 1000 new training samples. These samples are made up from pure compounds like those in Figure 3; however, in this figure A and C have been shifted in position (in other words, they are new compounds with new spectra). The cluster is still elliptical, although its size, center, and orientation have changed.

Figure 6. Two-wavelength spectra of three hypothetical compounds like those in Figure 3. There are 1000 samples of each compound. Unlike in Figure 3, however, error was permitted in the location of each compound, giving the three basically elliptical clusters shown (B and C at the top overlap).

Figure 7. Spectra of 1000 training samples, composed of the same compounds used to generate Figure 5. This time error was permitted in the locations (spectra) of the compounds A, B, and C (see Figure 6) and the shape of the resulting training cluster is irregular. The cluster corresponding to compound B was three times larger than those of A and C.

Figure 8. The frequency distribution of the points inside a cylinder containing the center of the cluster shown in Figure 7. The distribution is right-skewed.

Figure 9. The effect of skew on confidence-limit determinations. Symmetric limits (about the mean) are clearly inadequate for this distribution. The probability of an observation appearing at the (mean-a) and the (mean+a) is not the same.

Figure 10a. A hypothetical true population distribution F, containing all possible training samples.

Figure 10b. Constructing an empirical distribution (the training set) from the possible training samples (the unknown distribution F) in Figure 10a.

Figure 11a. Drawing a "bootstrap set" from the training set. This set has the same number of samples as the training set. The bootstrap-set samples are selected randomly and with replacement from the samples in the training set. The bootstrap-set samples appear as vertical-lined bars. The training-set samples appear (to provide a position reference) as crosshatch bars.

Figure 11b. The unknown distribution F is approximated by repeated drawing of randomly selected bootstrap-sample sets, of the same size as the training set, from the training set. In each bootstrap-sample set some training-set values are selected once, some more than once, and some are not selected at all. Again, the bootstrap-sample set appears as vertical-lined bars, and for reference, the unchanged training set appears as crosshatched bars.

Figure 11c. A third unique bootstrap-sample set (vertical-lined bars) drawn from the same training-set samples (crosshatched bars) used in Figures 11a and 11b.

Figure 12. The hypothetical distribution of parameters calculated for each bootstrap-sample set generated during the Monte-Carlo process depicted in Figures 11a, 11b, and 11c.

Figure 13. The percentage bias (solid line) and RSD (dotted line) of the BEAST distance estimator as a function of the training-set size (N) (B=1000 bootstrap replications, D=2 wavelengths (dimensions), R=0.001, the radius of the hypercylinder).

Figure 14. The dependence of the relative standard deviation of the BEAST estimator on the number of bootstrap replications (B) of the training set. The solid line corresponds to a sample monitored at five wavelengths, the dotted line to a sample monitored at three wavelengths, and the dashed line to a sample monitored at only one wavelength (N=50 training samples, R=0.001, the radius of the hypercylinder).

Figure 15. The percentage bias of the BEAST estimator as a function of the number of bootstrap replications (B) of the training set (N=150 training samples, D=5 wavelengths (dimensions), R=0.005, the radius of the hypercylinder).

Figure 16. The relative standard deviation (dotted line) and bias (solid line) of the BEAST estimator as a function of the radius (R) of the hypercylinder (B=50 bootstrap replications, N=100 training samples, D=5 wavelengths (dimensions)). The high RSD values observed are the result of using only 50 bootstrap replications.

Figure 17. The relative standard deviation (dotted line) and bias (solid line) of the BEAST estimator as a function of the number of wavelengths (D) used to monitor the sample (B=10000 bootstrap replications, N=50 training samples). The hypercylinder radius was set at R=0.001 for two and three wavelengths, and at R=0.005 for five wavelengths.

Figure 18. A cross-section of the benzoic acid derivatives training set used in Tables II and III. This cross-section has been projected on a wavelength axis, and is right-skewed.
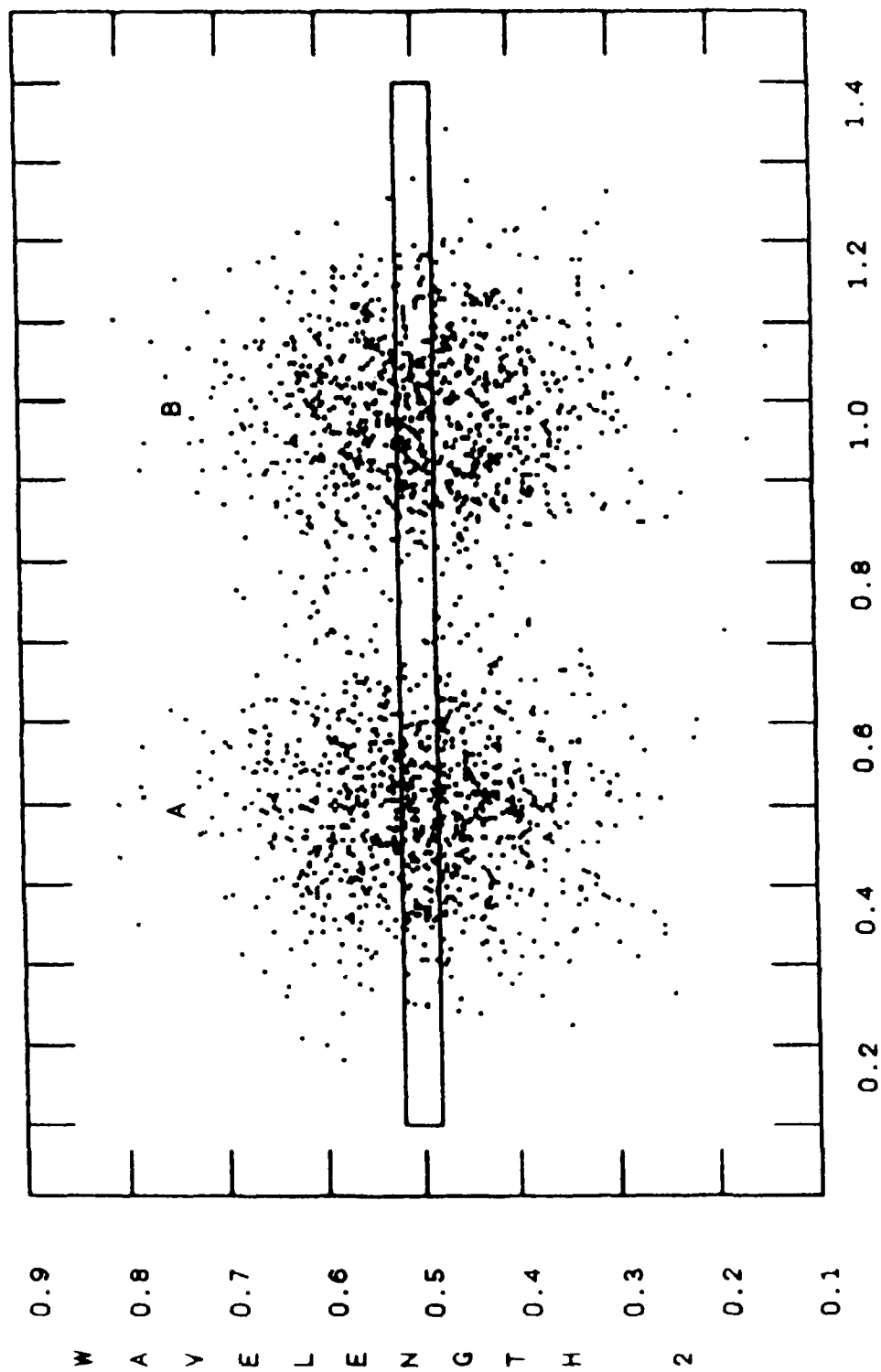
Fig. 1
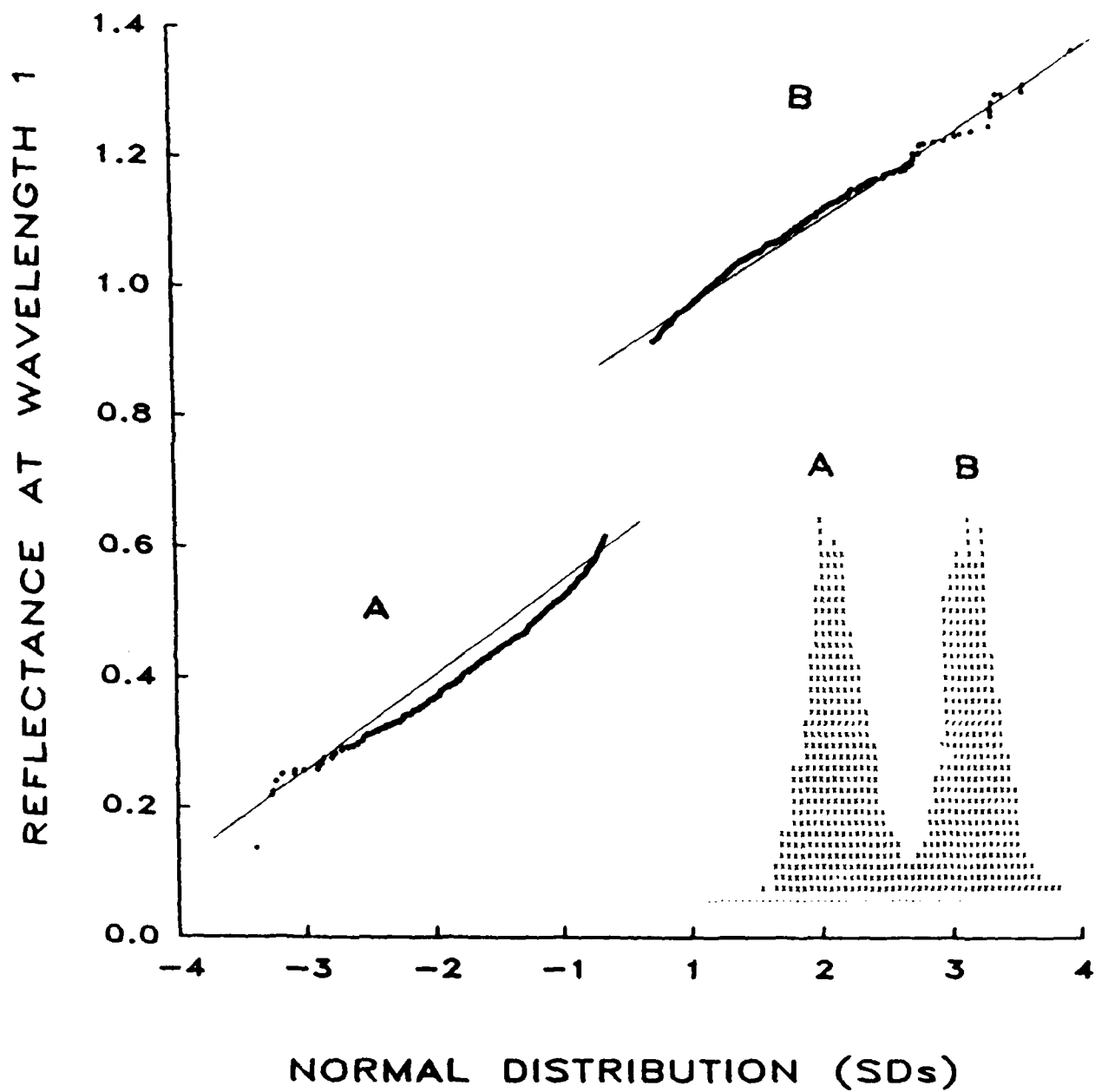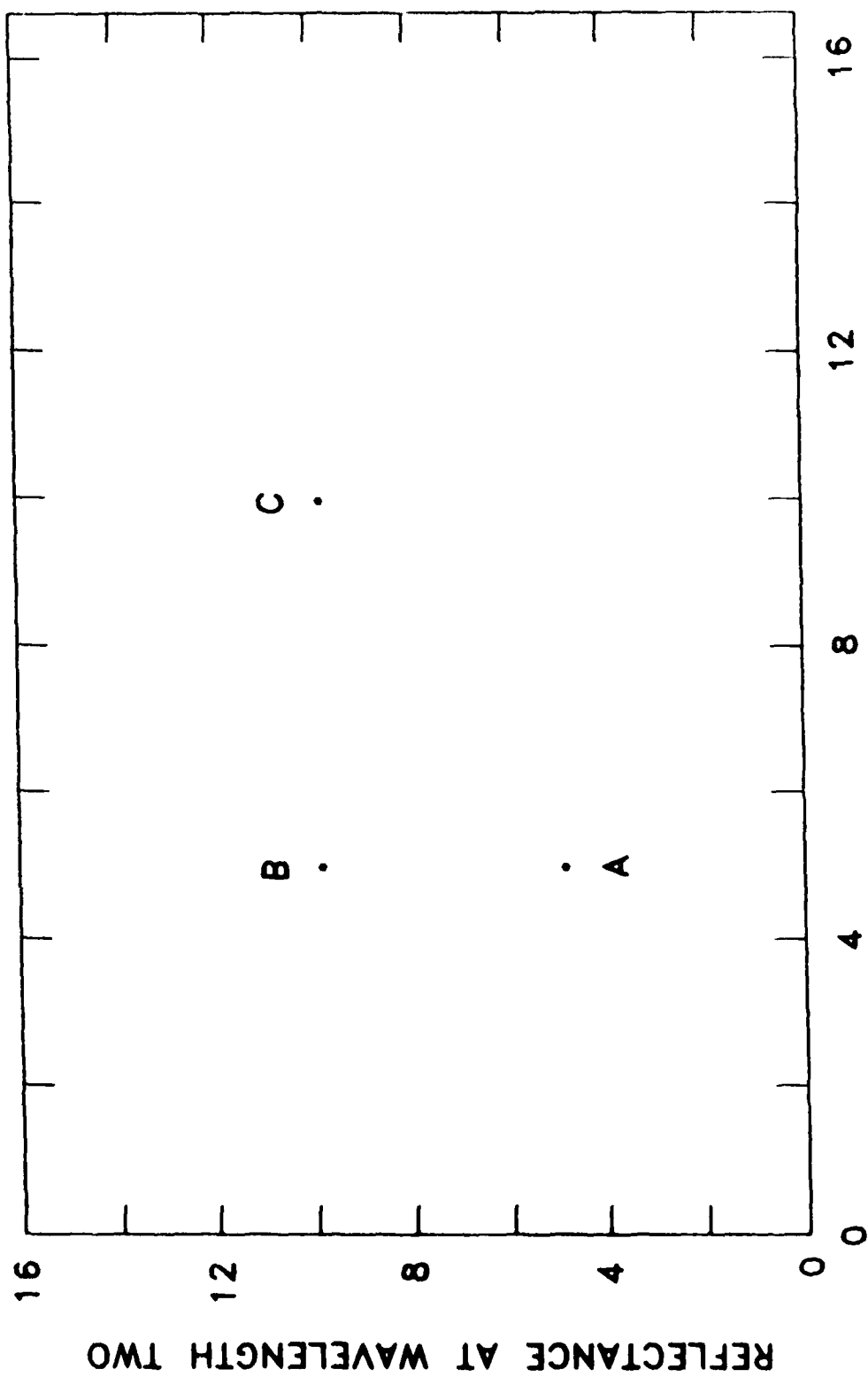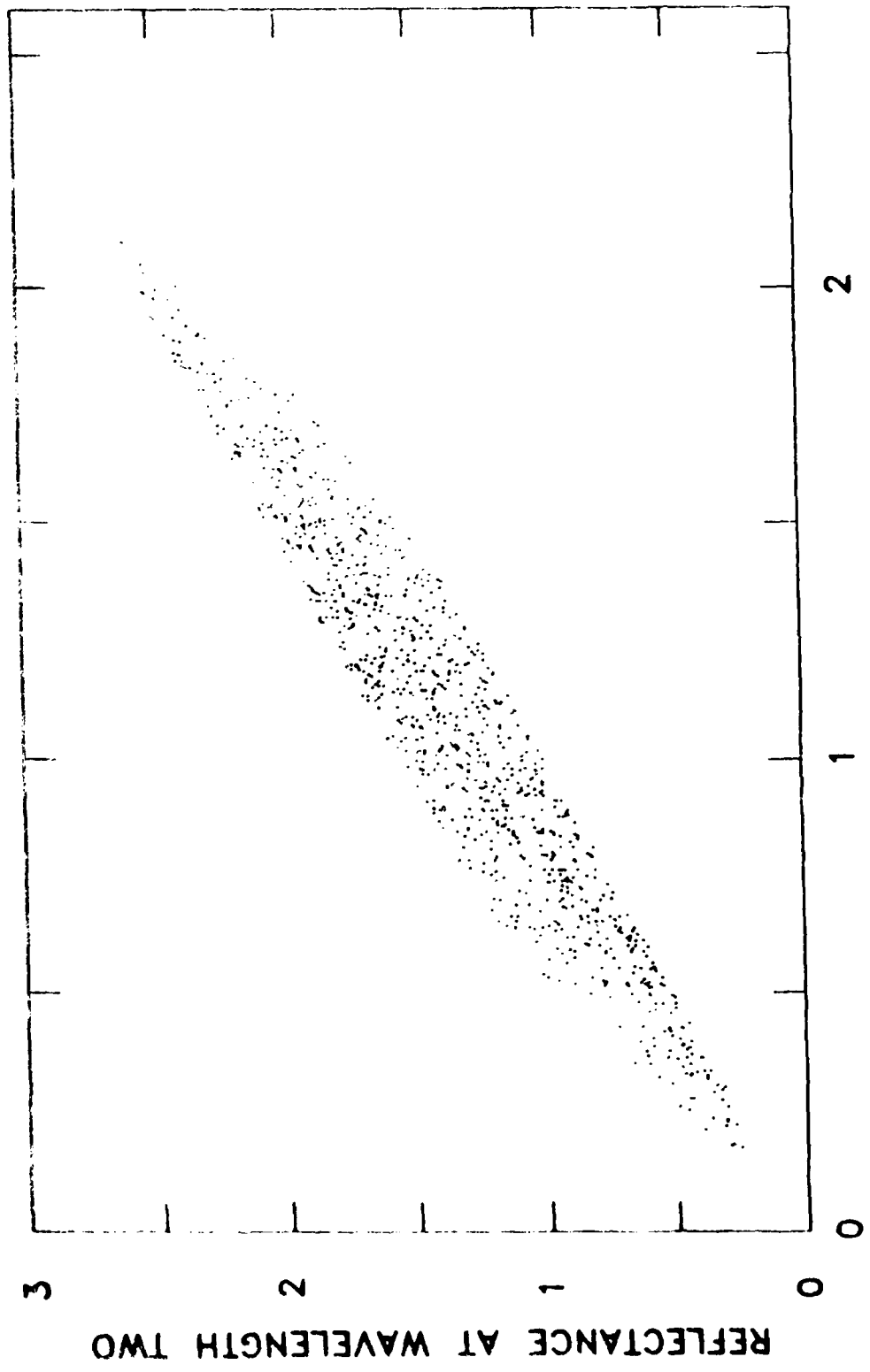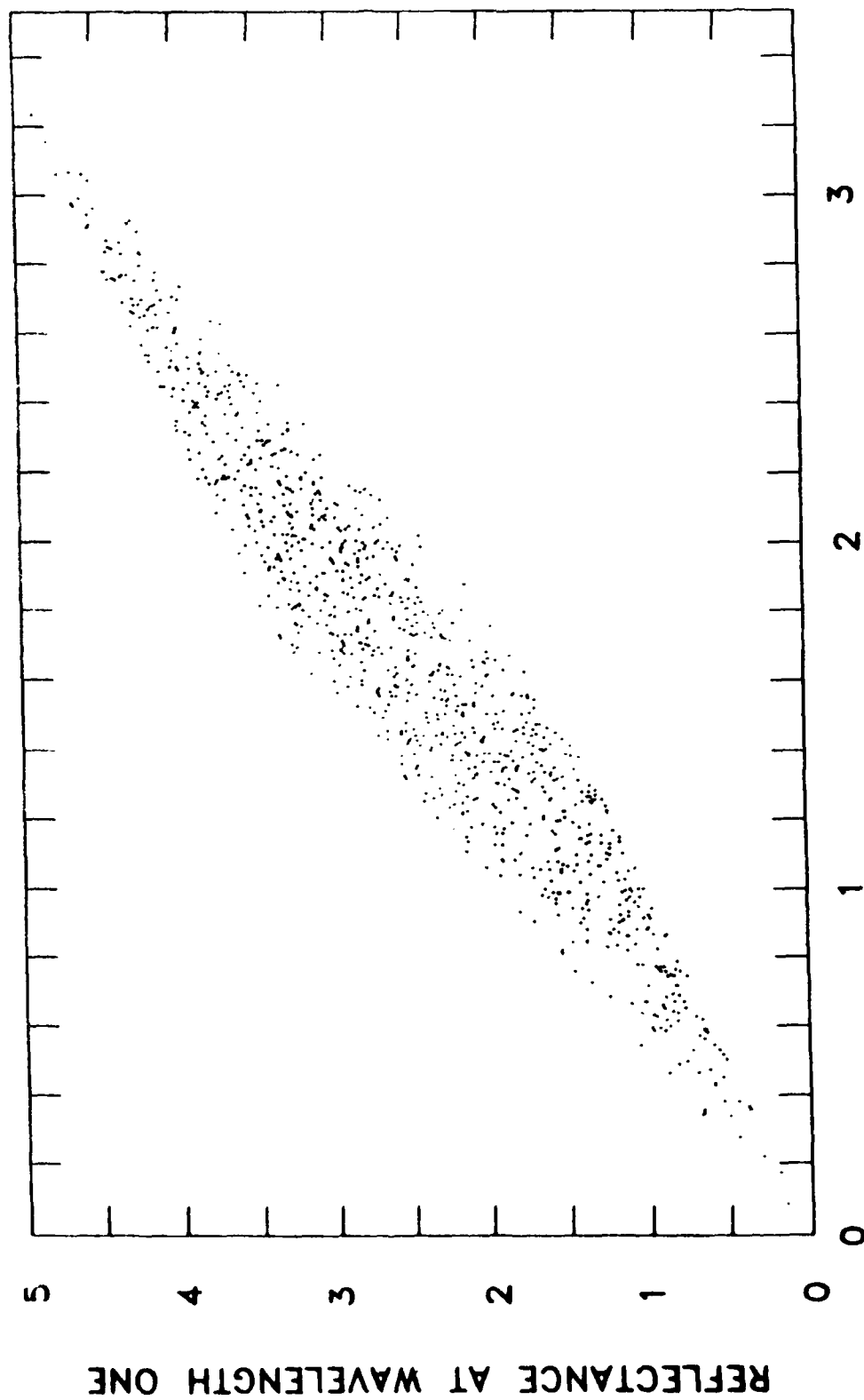
WAVELENGTH 1

W A V E L E N G T H 2

Fig. 2

Fig. 3

REFLECTANCE AT WAVELENGTH ONE

REFLECTANCE AT WAVELENGTH TWO

Fig. 1

REFLECTANCE AT WAVELENGTH ONE

REFLECTANCE AT WAVELENGTH TWO

REFLECTANCE AT WAVELENGTH TWO

REFLECTANCE AT WAVELENGTH ONE

Fig 5

Fig. 6

REFLECTANCE AT WAVELENGTH ONE

REFLECTANCE AT WAVELENGTH TWO

REFLECTANCE AT WAVELENGTH ONE

REFLECTANCE AT WAVELENGTH TWO

Fig. 2

Fig. 8

FREQUENCY

17.5 15.0 12.5 10.0 7.5 5.0 2.5 0.0

REFLECTANCE

1   2   3

Fig. 9

DISTANCE

FREQUENCY

Fig. 2a

Fig. 10 A

DISTANCE

FREQUENCY

Fig 1a

DISTANCE

FREQUENCY

Fig. 4

DISTANCE

FREQUENCY

Fig. 10

FREQUENCY

DISTANCE

FREQUENCY

DISTANCE

Fig. 2.

Fig. 13

BIAS AND RSD (%)

TRAINING-SET SIZE (N)

NUMBER OF BOOTSTRAP REPLICATIONS (B)

RELATIVE STD. DEVIATION (%)

BIAS (%)
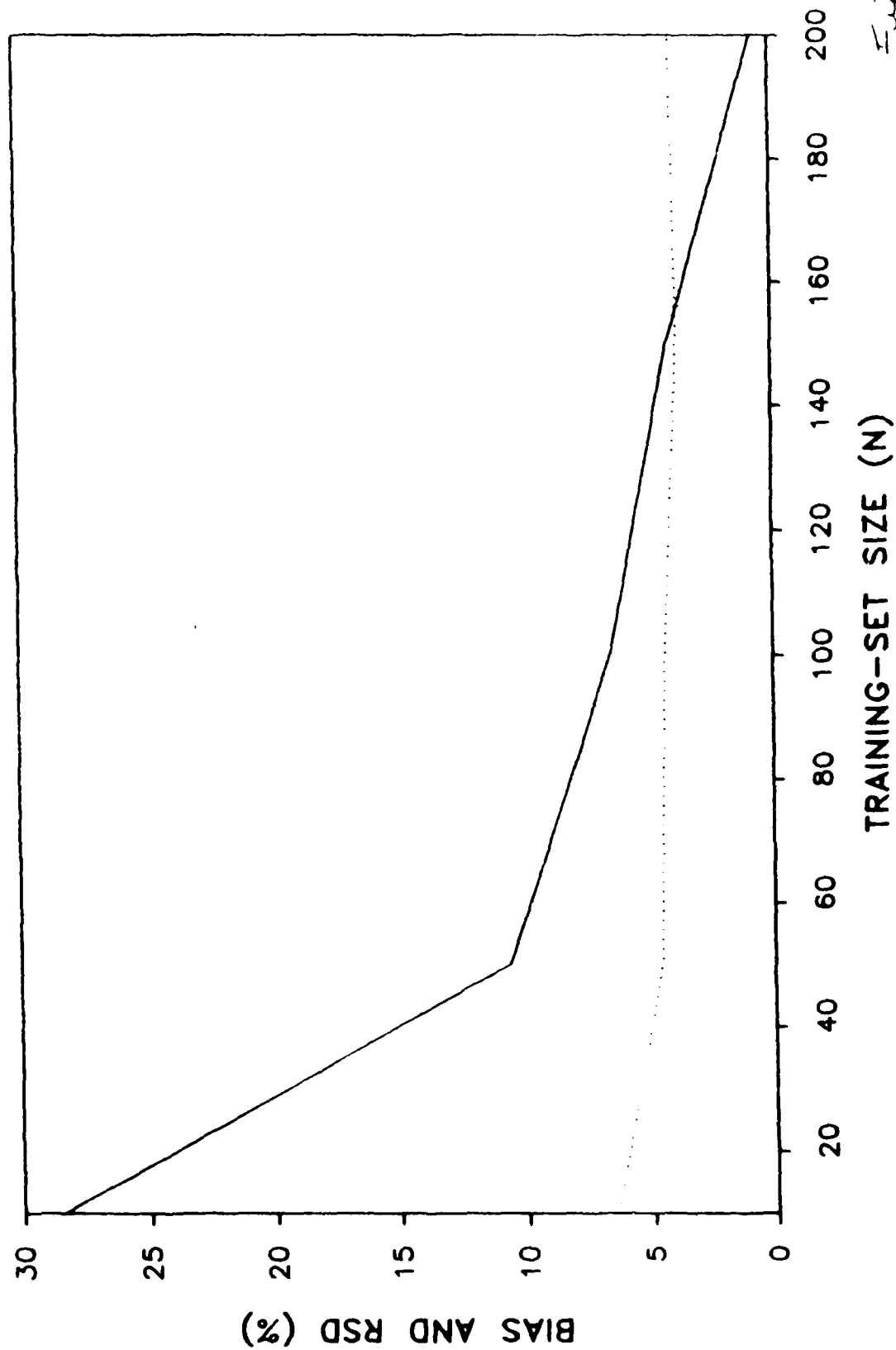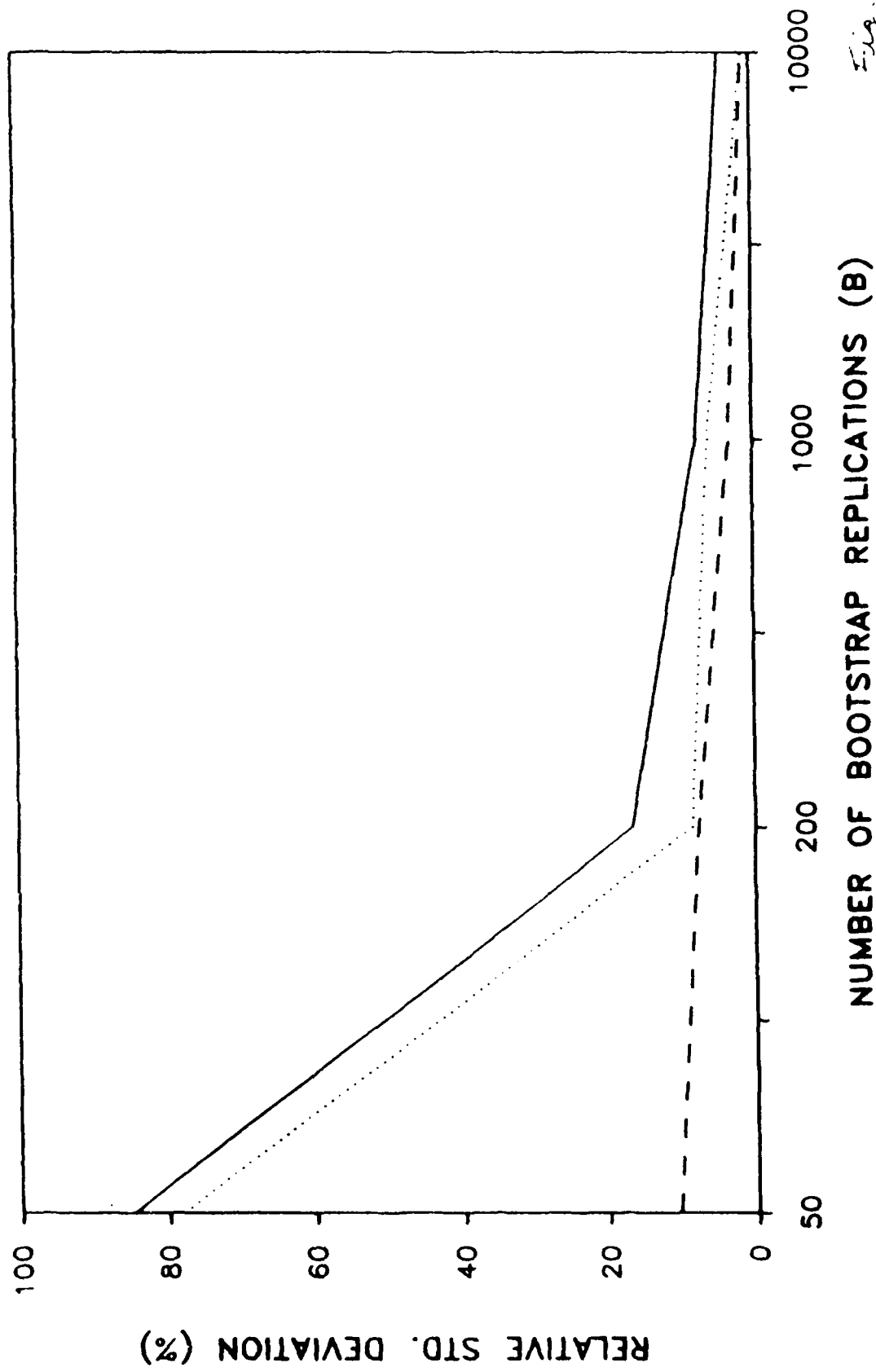
NUMBER OF BOOTSTRAP REPLICATIONS (B)

Fig. 10

RADIUS OF HYPERCYLINDER

BIAS AND RSD (%)

NUMBER OF WAVELENGTHS (D)

BIAS AND RSD (%)
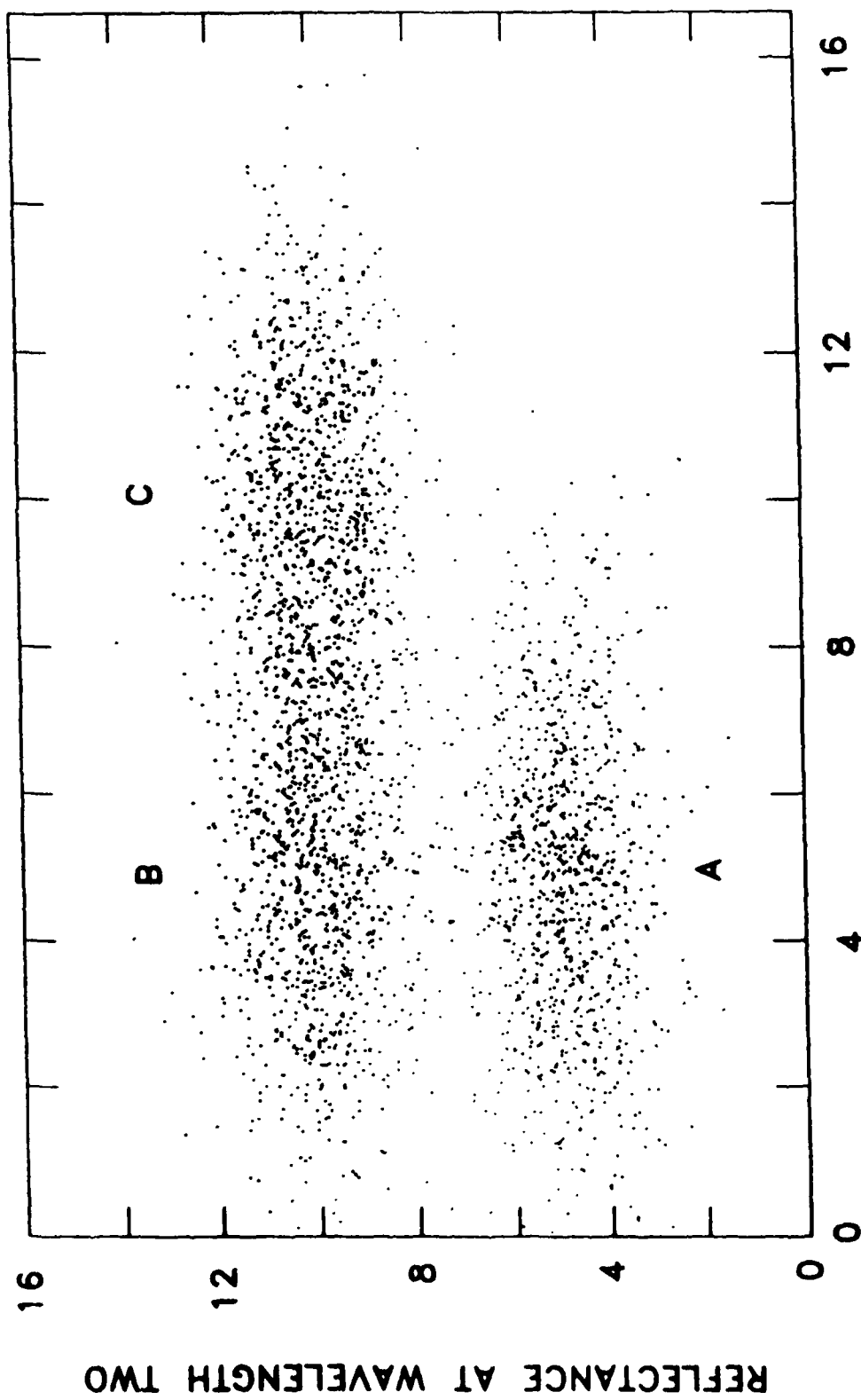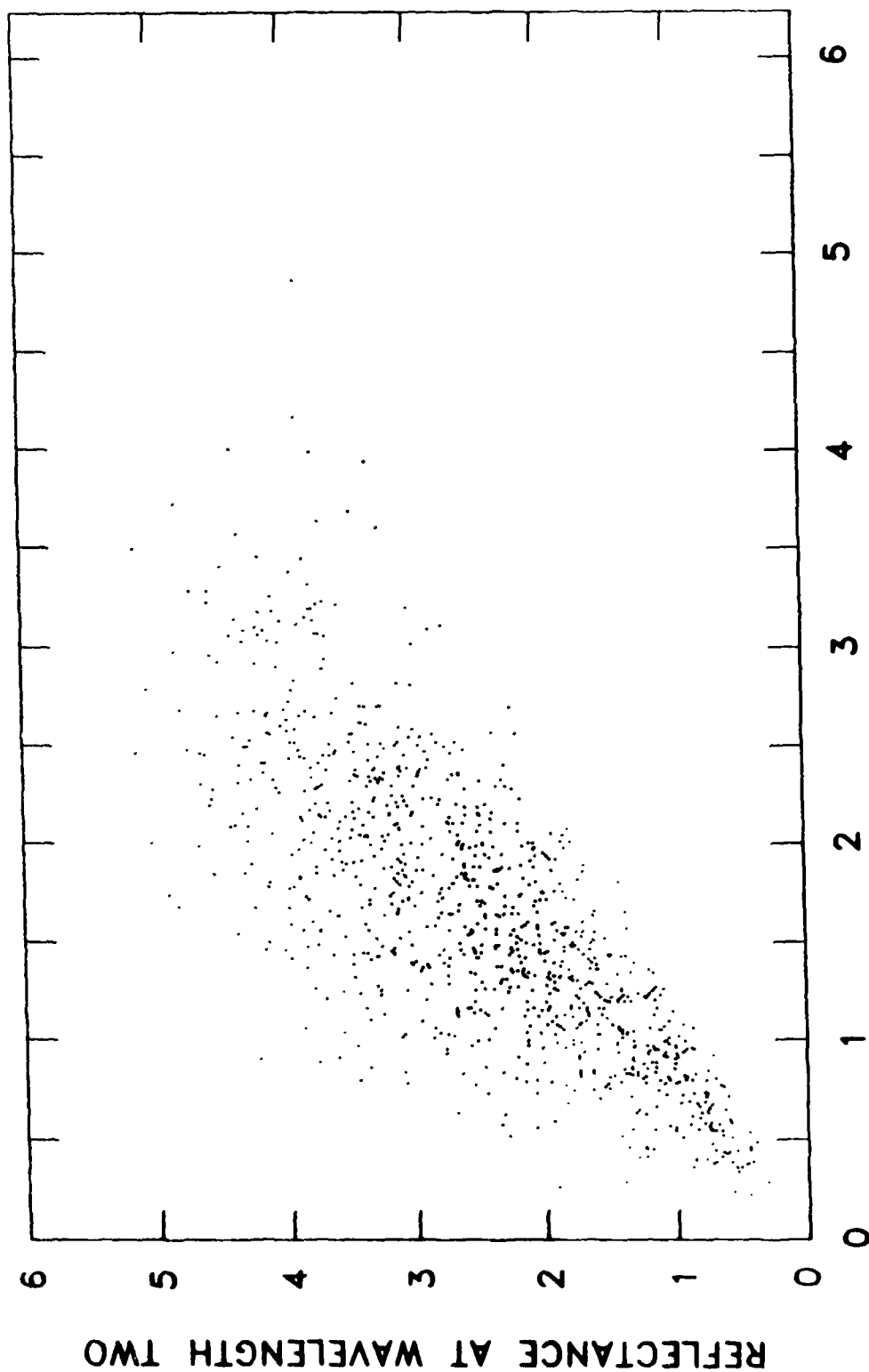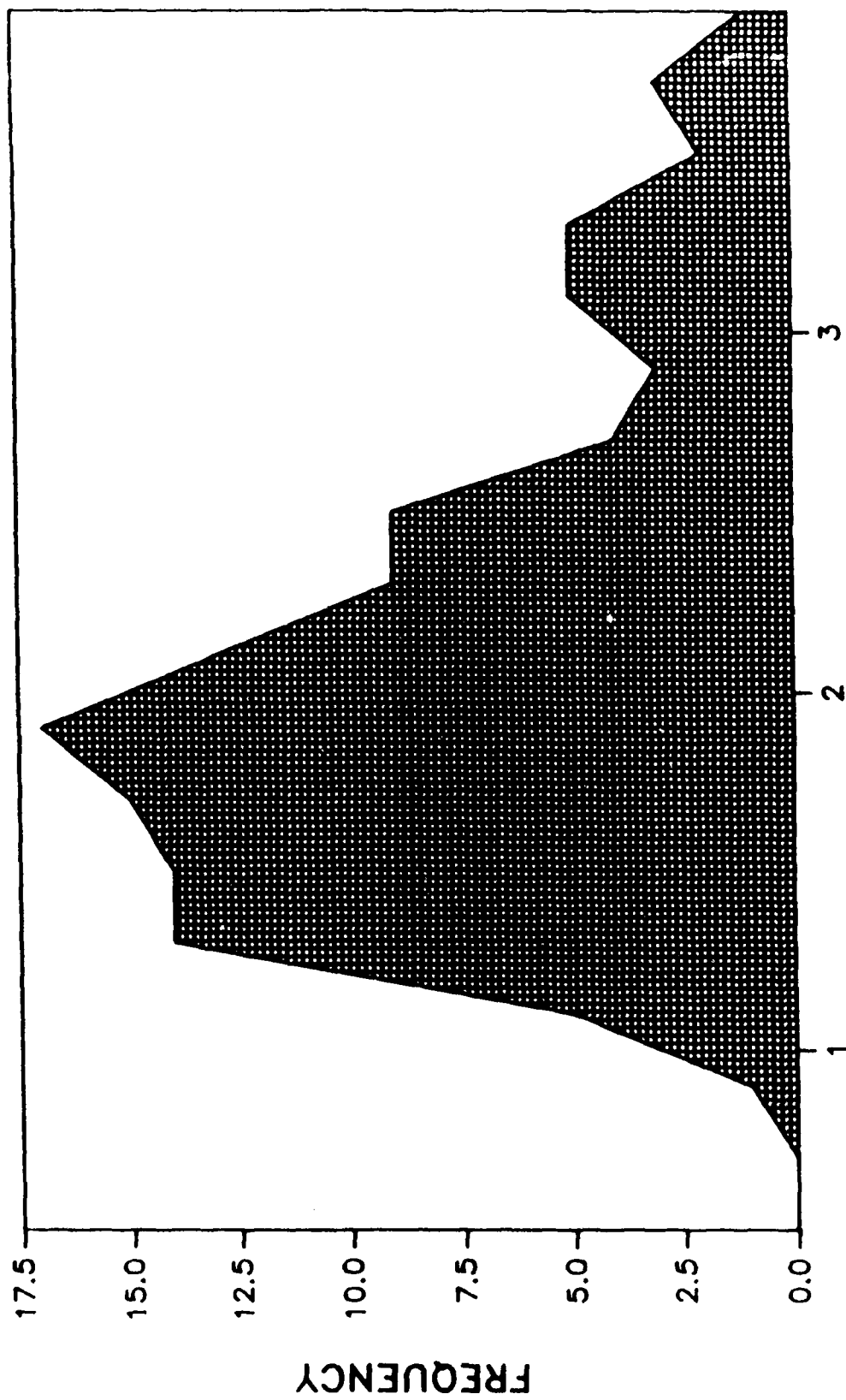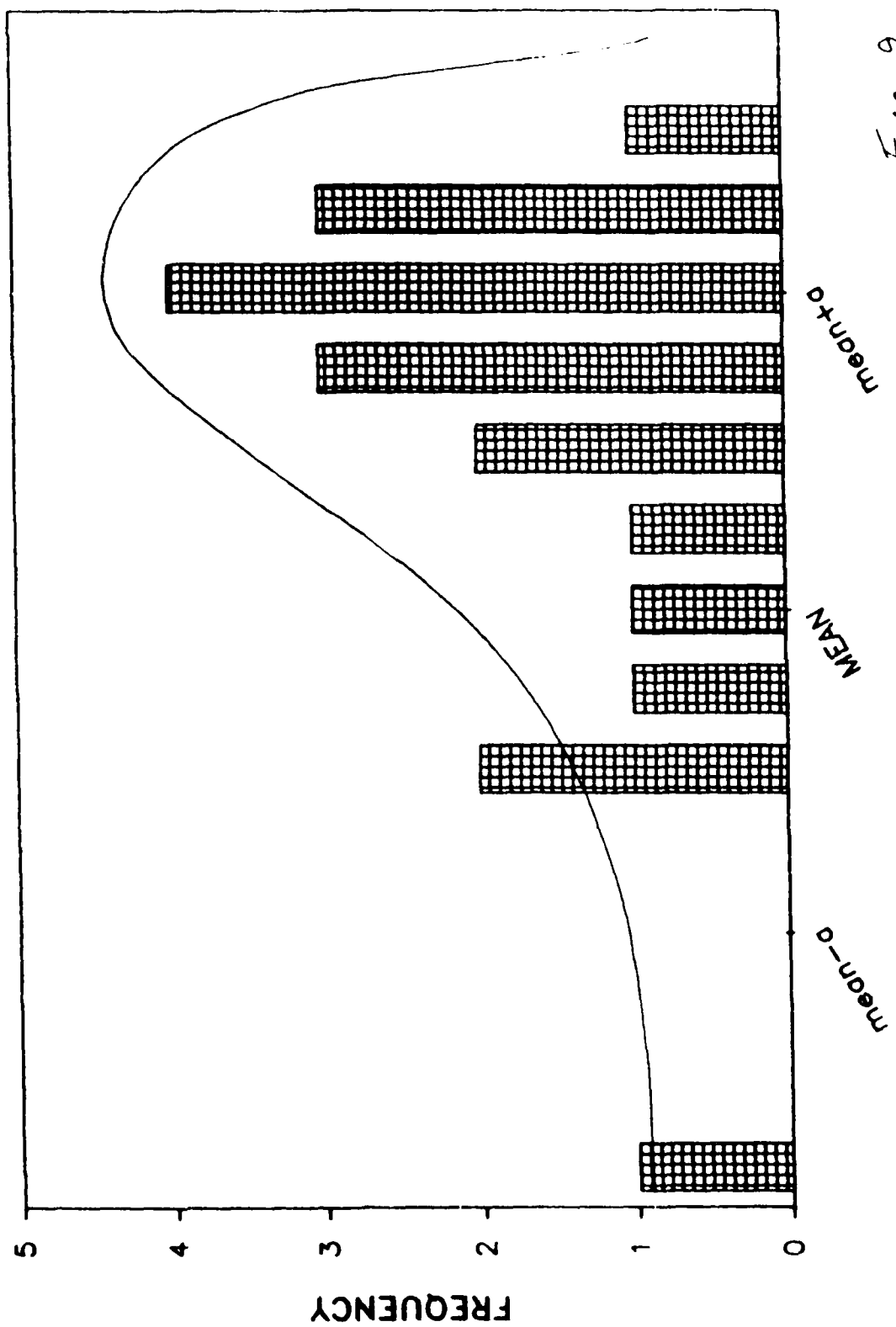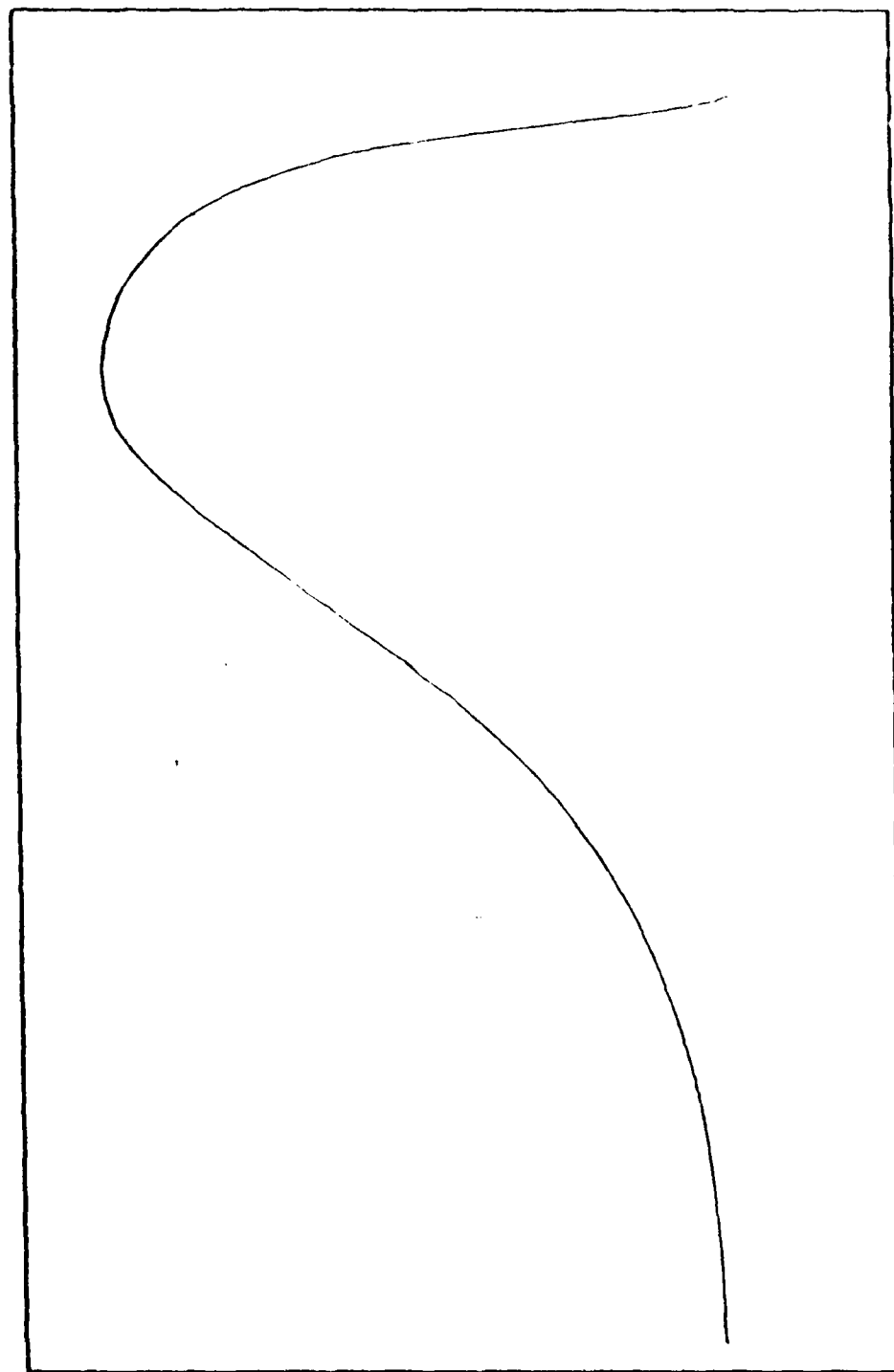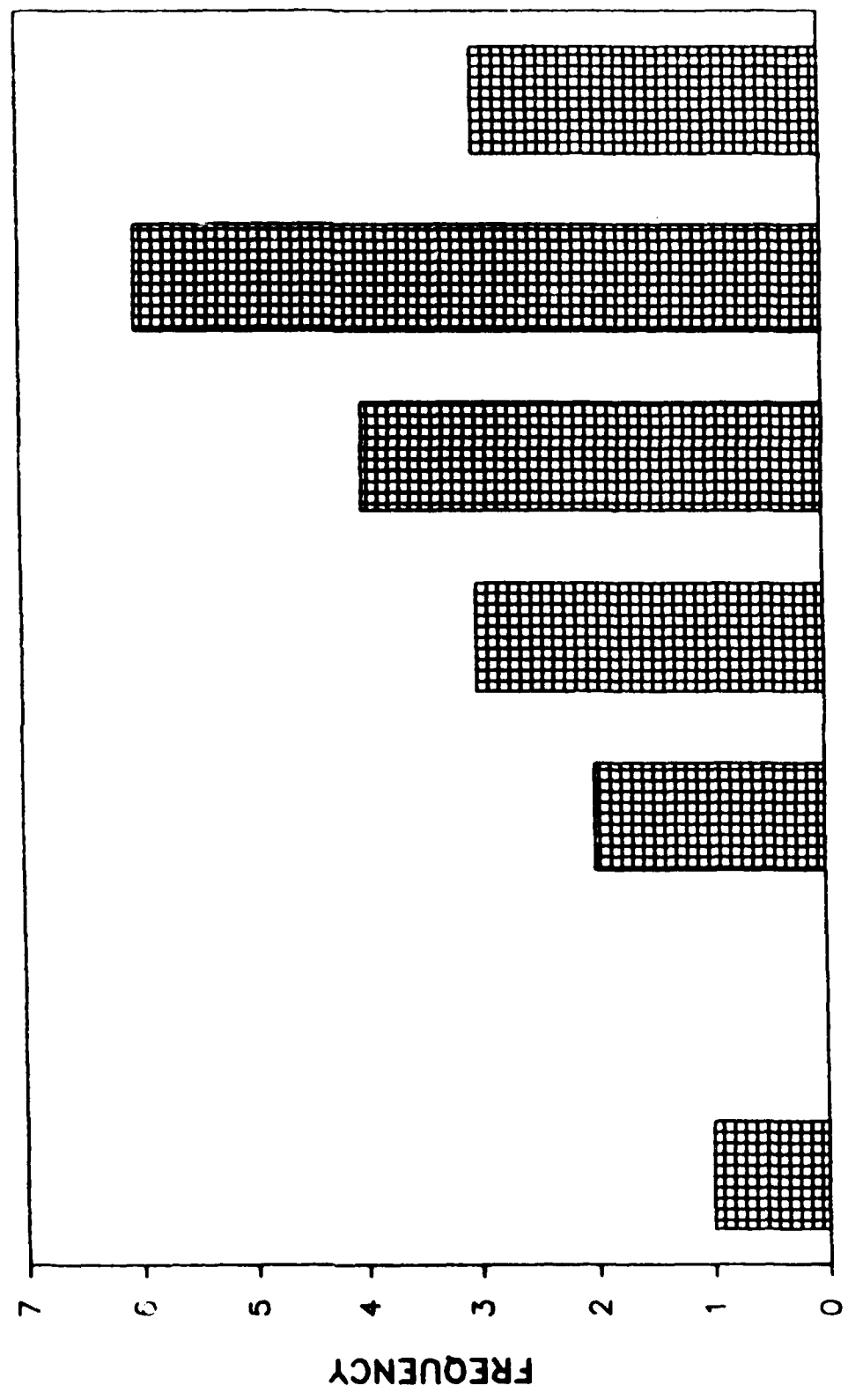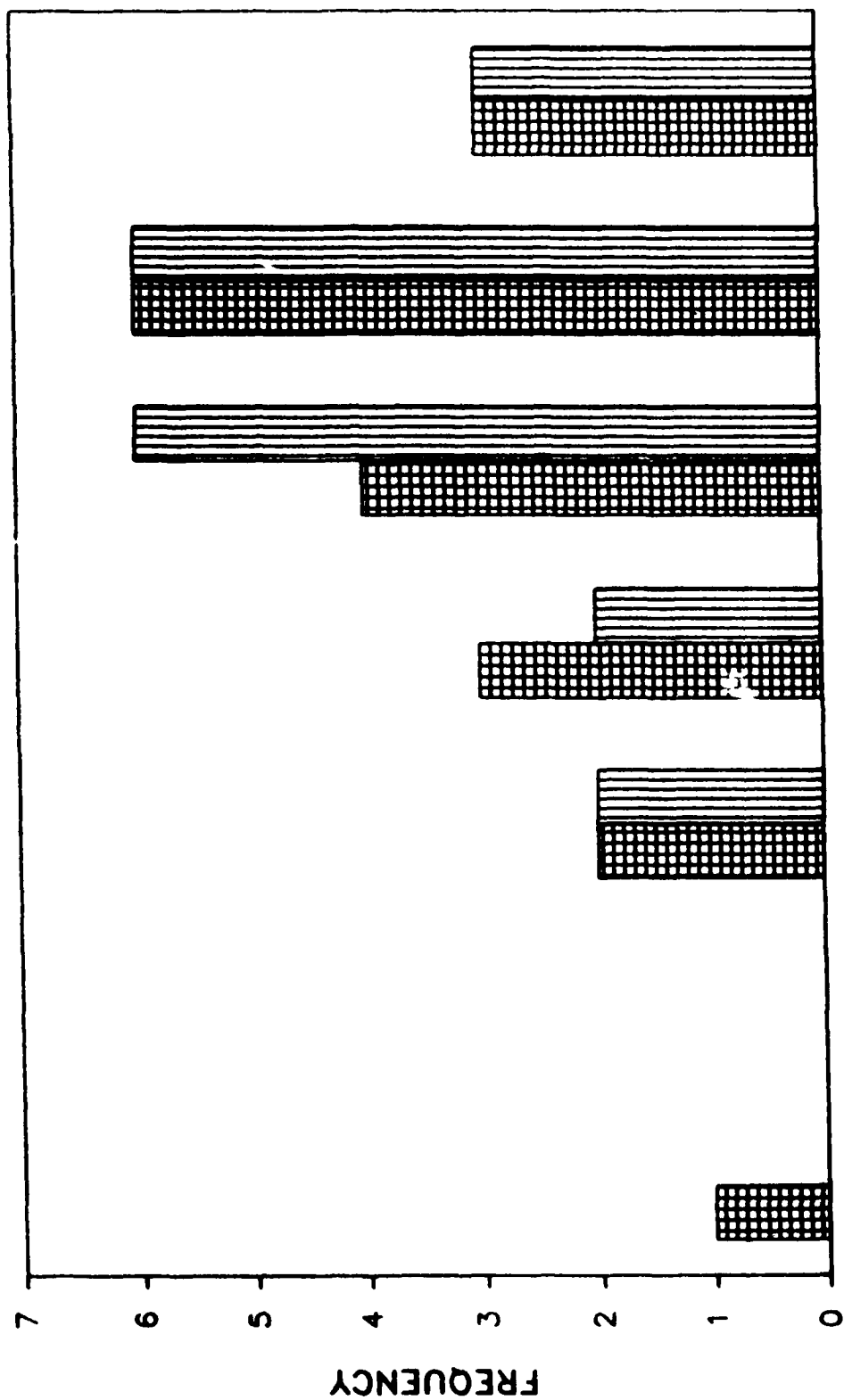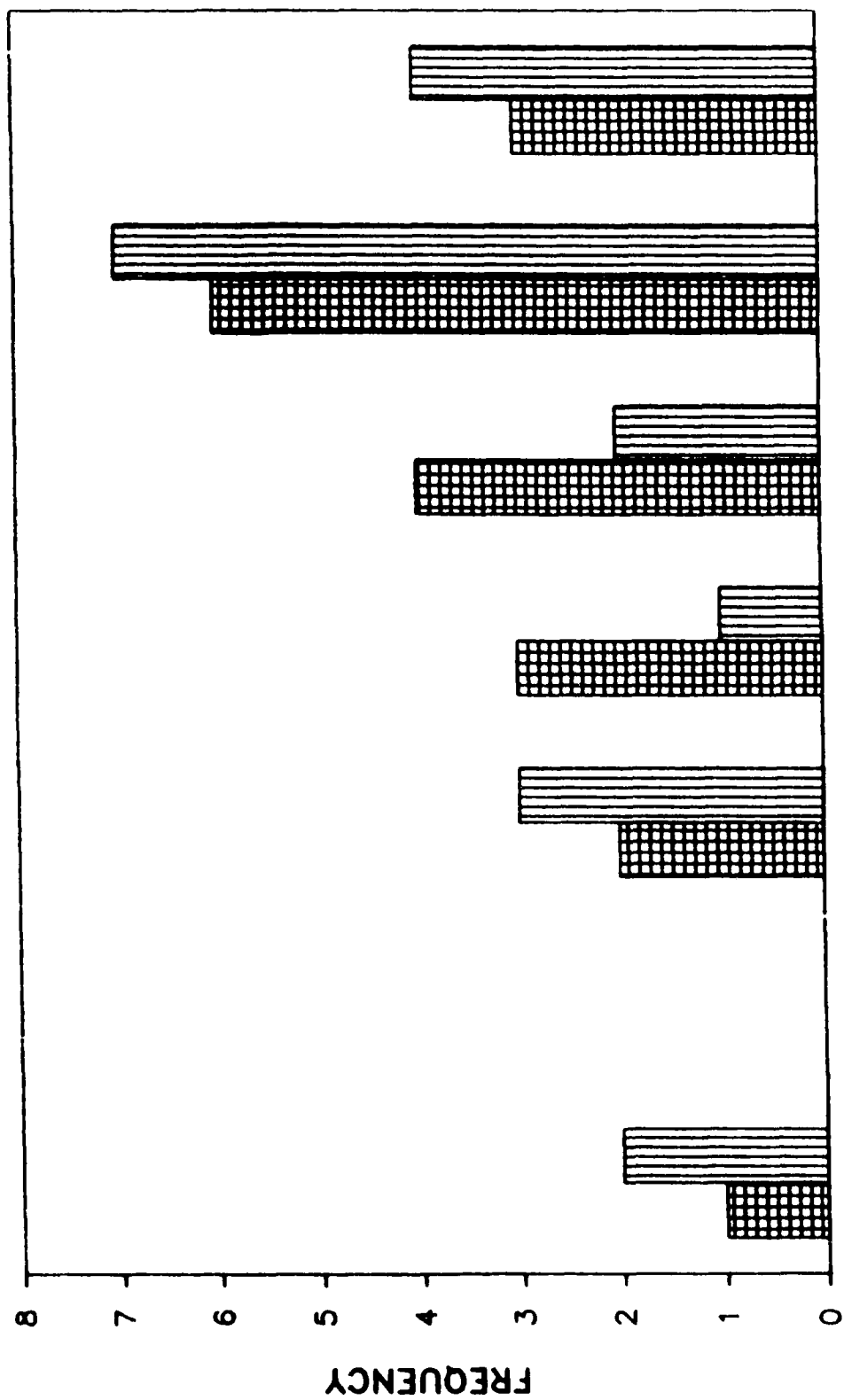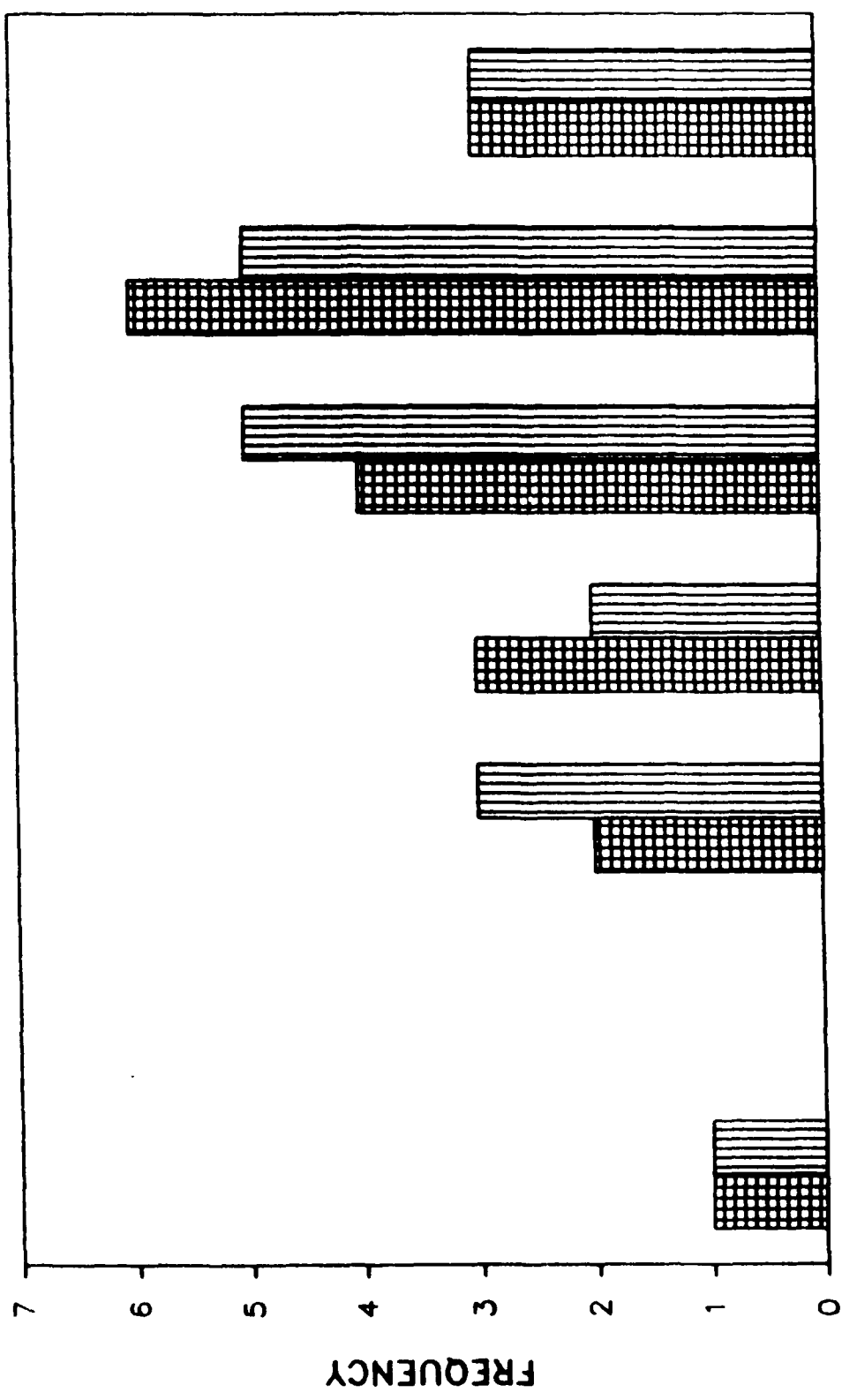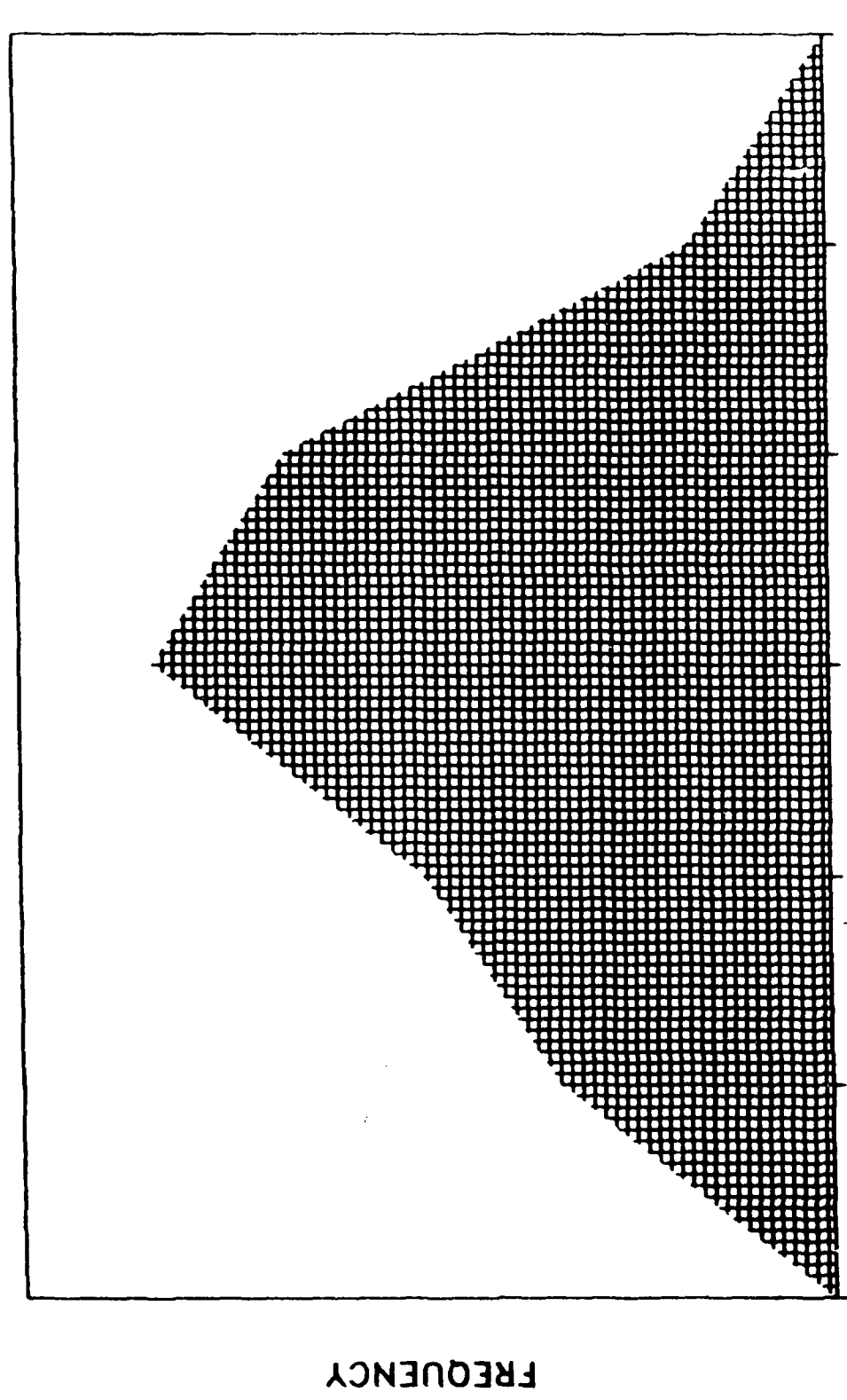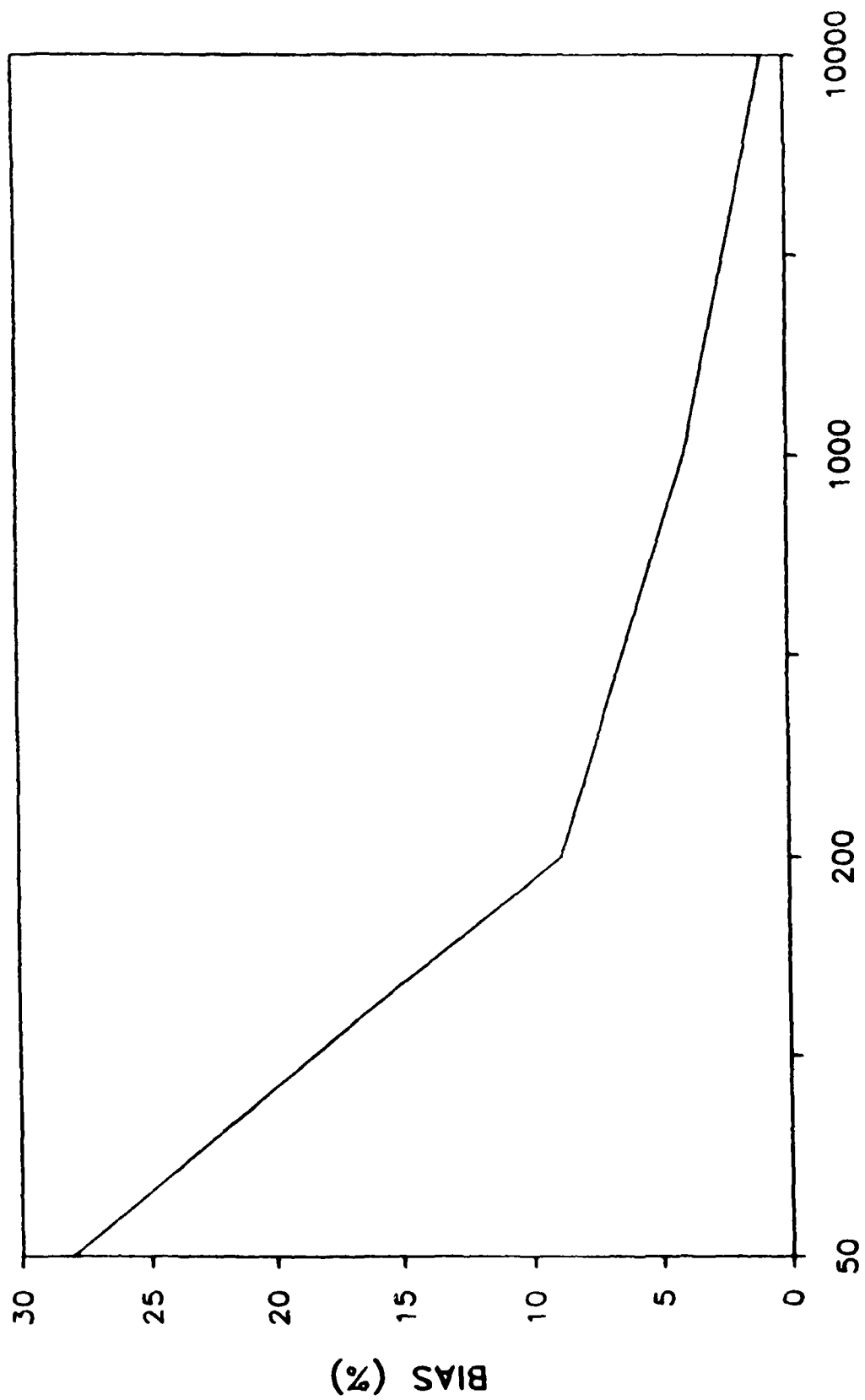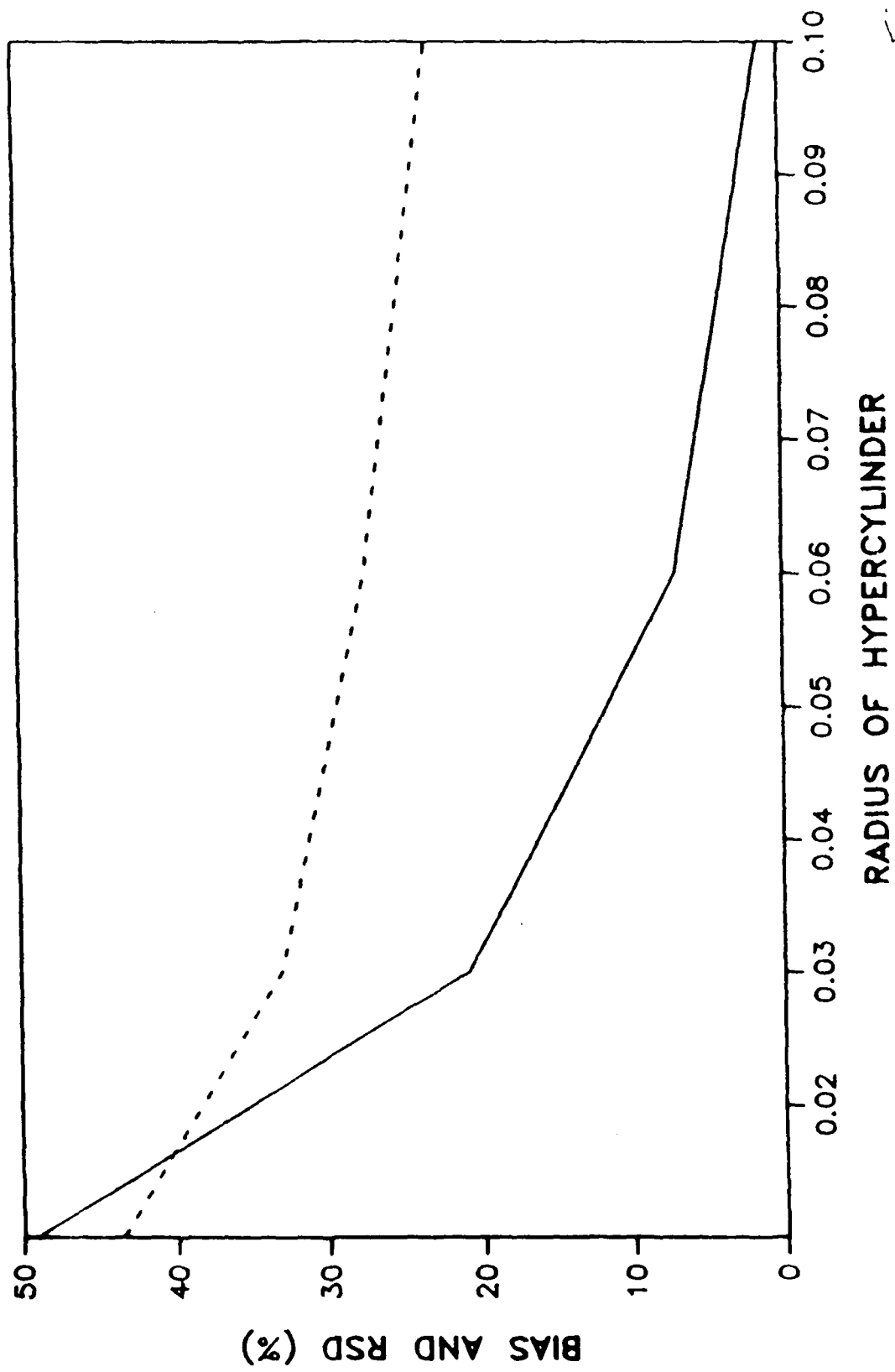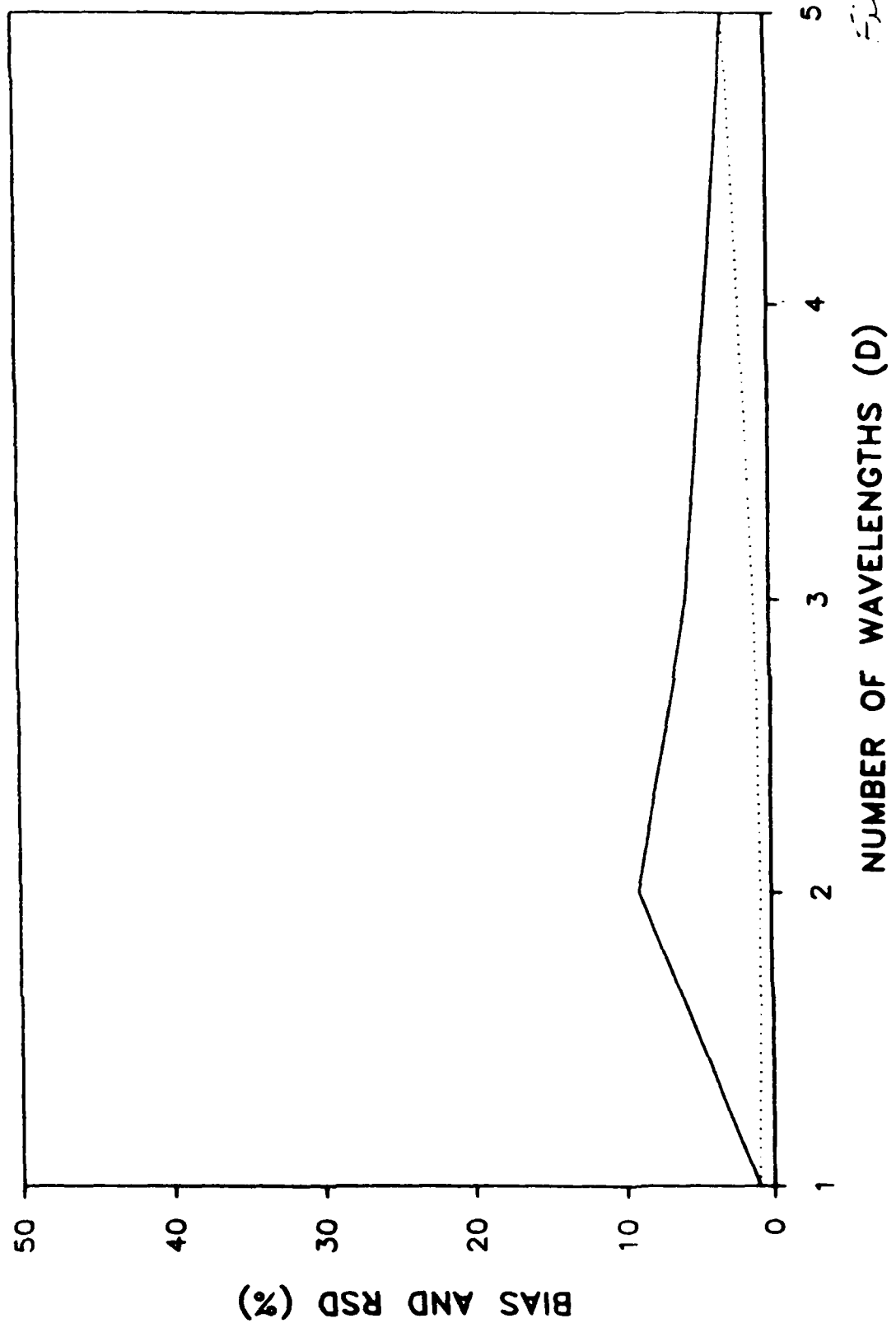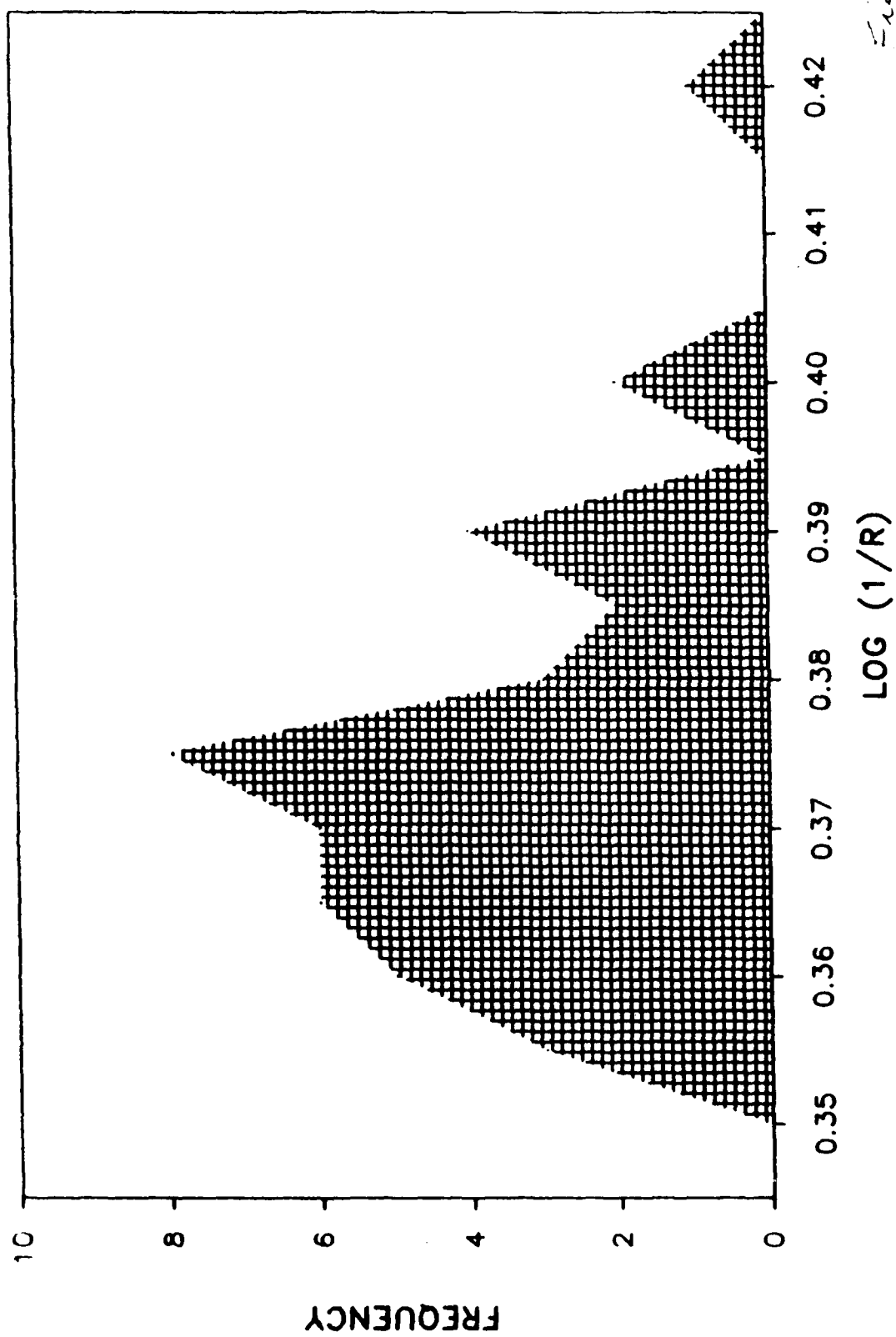
FIG. 2

Special defined operations:

$M(x)$ median of x (x is a set, vector, 1- or 2-D array)

$R(f(x))$ roots of f(x) by trapezoidal interpolation

r random number on 0<x<1, Monte Carlo integration of continuous uniform distribution

$\kappa(T)$ creates bootstrap distribution B for training set T, and finds the center C of the distribution

$\psi(T,B,X,C)$ finds BEAST distance from center C of training set T to new spectrum X using probability determined with bootstrap distribution B

$[x]$ greatest integer function of scalar, set, vector, or 1- or 2-D array

$\phi(x)$ $1/(2\pi)^{1/2} \int_{-\infty}^{x} e^{-t^2/2} \, dt$,   area from $-\infty$ to x.

$\phi^{-1}(x)$ inverse of above; i.e., given area, find x

-1-

$O(x)$ ordered elements of x (x is a set, vector, 1- or 2-D array

$=$ equals, or "is replaced by" when the same variable appears on both sides of =

| "such that" qualifier on a variable, e.g., $\{x | 0 < x < 1\}$ specifies the range of possible values for x

$|x|$ the absolute value of x (x is a scalar)

Scalars:

n training-set size, i.e., number of samples

d number of wavelengths

m number of training-set replications comprising bootstrap distribution (user determined)

σ BEAST standard deviation (SD), average of upper and lower confidence limits producing a symmetric distance

σ_c error-adjusted BEAST SD, asymmetric value produced

-2-

using only one confidence limit

r_h hypercylinder radius (user-determined)

δ skew sensitivity (user-determined)

n_h number of spectral points falling inside a hypercylinder

l lower confidence-limit index (index is a position
in an ordered array that expresses the value of an
integral from the end of the array to the index)

u upper confidence-limit index

α contour level specified by $\Phi(-\sigma)$, used to
determine if test spectrum is inside or outside a
cluster

z_α $\Phi^{-1}(\alpha)$

$S_{(O2)}$ Euclidean distance from bootstrap-distribution
center C to new spectral point X

$S_{(COR)}$ Euclidean distance from training-set center $C_{(T)}$
to bootstrap-distribution center C

$S_{(C2R)}$ Euclidean distance from training-set center $C_{(T)}$

to the new spectral point X

$S_{(CUB)}$ 1/2 the total length of the sides of a triangle specifying a particular plane in hyperspace

A_c area of a triangle whose vertices specify a particular plane in hyperspace

$S_{(CR)}$ Euclidean distance from training-set center $C_{(T)}$ to hyperline connecting C to X

$S_{(CP)}$ Euclidean distance $S_{(COR)}$ projected on the hyperline connecting C to X

z_e index for error adjustment in S_q

z_o $\phi^{-1}(z_e/n_h)$

Matrices, vectors, and arrays:

$B = (b_{ij})_{m,d}$ m by d bootstrap distribution

$C = (c_j)_d$ center of the bootstrap distribution B

$P = (p_{ij})_{m,n}$ training-set sample numbers selected for bootstrap-sample sets used to calculate the

bootstrap distribution

$B_{(\bullet)} = (b_{(\bullet)ij})_{n,d}$ bootstrap sample set used to calculate single rows of B

$T = (t_{ij})_{n,d}$ training-set sample spectra

$X = (x_j)_d$ test-sample spectrum

$S_{(OR)} = (s_{(OR)i})_m$ Euclidean distances from each element of B to C

$S_{(2R)} = (s_{(2R)i})_m$ Euclidean distances from each element of B to X

$S_{(UB)} = (s_{(UB)i})_m$ 1/2 total length of triangle sides formed by planes in hyperspace connecting X, C, and the rows of B

$A = (a_i)_m$ areas of triangles formed by planes in hyperspace connecting X, C, and the rows of B

$S_{(R)} = (s_{(R)i})_m$ radial Euclidean distances from the rows of the bootstrap distribution B to the hyperline connecting X to C

$S_{(P)} = (s_{(P)i})_m$ Euclidean distances from C to the rows of B projected on the hyperline connecting C to X

$S_{(q)} = (s_{(q)i})_{nh}$ ordered n_h elements of $\{s_{(p)i}|r_h < s_{(R)i}\}$

$C_{(T)} = (c_{(T)j})_d$ center of training set T by $M(t_{ij})$

$F = (f_i)_{nh}$ n_h elements of $S_{(q)}$ corrected using $C_{(T)}$

$I_{(N)} = (i_{(N)i})_{nh}$ n_h independent variables of the set $\{1, 2, 3, \ldots\}$

                                        paired with F to locate the root of F

APPENDIX I.

The Quantile BEAST

The Quantile BEAST is basically an experimental clustering technique for exploring multivariate data distributions. A number of different variations in the details of the implementation still produce a method that is consistent with the description presented thus far. As is the case with many problems, there is more than one route to the same solution. One path to determining the BEAST distance is presented here.

In near-infrared spectral data analysis, virtually all implementations of the BEAST begin with the collection of a training set of samples. The training set consists of spectral data values (e.g., absorbance, $\log(1/R)$, etc.) recorded at d wavelengths for n training samples. The resulting data are represented by a two-dimensional n-by-d matrix (or array) T.

The BEAST itself is composed of two operations:

1. The bootstrap distribution is created from the training set by an operation $\kappa(T)$. This bootstrap distribution forms the basis for calculating directional probabilities, and is calculated only once for each training set. $\kappa(T)$ provides the bootstrap distribution B for the training set as well as the center C (groupmean) of the bootstrap distribution.

2. The operation $\psi(T,B,X,C)$ calculates σ and σ_c (the BEAST standard deviation, or SD) using the training set T, the bootstrap distribution B and center C (from step 1 above), and the test

-7-

sample's spectrum X. The Euclidean distance from C to X is scaled by σ or σ_c to give the distance to X in BEAST SDs.

Once the training set has been assembled $\kappa(T)$ can be calculated. Random selections are made from T by filling P with the training-set sample numbers to be used in the bootstrap sample sets $B_{(s)}$,

$$P = p_{ij} = r \tag{1}$$

and then the values in P are scaled to the training-set size n:

$$P = [(n-1)P + 1]. \tag{2}$$

A bootstrap sample $B_{(s)}$ is then created for each row i of B by

$$B_{(s)} = t_{kj} \tag{3}$$

where k are the elements of the ith rows of P. The ith row of B is actually filled by the center (groupmean) of the bootstrap sample

$$b_j = \sum_{i=1}^{n} b_{(s)i,j} / n \tag{4}$$

and the center of the bootstrap distribution is

$$c_j = \sum_{i=1}^{m} b_{ij} / m \tag{5}$$

At this point $\kappa(T)$ is complete for the training set, and the analysis of actual test samples can begin.

The calculation of the BEAST distance using $\psi(T,B,X,C)$ now requires only a test spectrum X obtained from a sample of interest by scanning the sample at d wavelengths. This calculation involves finding the hyperline connecting C and X, and determining the probability of X belonging to T based upon the number of points (rows of B) within a certain distance r_h

-8-

of the hyperline (in effect, taking the points or rows of **B** that fall within a hypercylinder of radius r_h). This implementation of the BEAST proceeds toward the point-density of the hypercylinder by forming planes connecting **X**, **C**, and the rows of **B**. The use of a series of planes allows a complex structure like a hypercylinder in d-dimensional hyperspace to be represented in a simple manner regardless of the number of spatial dimensions (in fact, the number of planes is completely independent of the spatial dimension).

The three points that specify a plane in space also specify a triangle whose sides are readily determined:

$$S_{(02)} = \left(\sum_{j=1}^{d} (x_j - c_j)^2 \right)^{1/2} \tag{6}$$

$$s_{(OR)i} = \left(\sum_{j=1}^{d} (b_{ij} - c_j)^2 \right)^{1/2} \tag{7}$$

$$s_{(2R)i} = \left(\sum_{j=1}^{d} (b_{ij} - x_j)^2 \right)^{1/2} \tag{8}$$

Once this series of triangles has been formed, finding the rows of **B** that fall inside the hypercylinder is a straightforward procedure.

$$s_{(UB)i} = (S_{(02)} + s_{(OR)i} + s_{(2R)i}) / 2 \tag{9}$$

$$a_i = (s_{(UB)i} (s_{(UB)i} - S_{(02)})(s_{(UB)i} - s_{(OR)i})(s_{(UB)i} - s_{(2R)i}))^{1/2} \tag{10}$$

$$s_{(R)i} = 2(a_i) / S_{(02)} \tag{11}$$

$$s_{(p)i} = (s_{(OR)i}^2 - s_{(R)i}^2)^{1/2} \tag{12}$$

The elements of $S_{(p)}$ are Euclidean distances, from the center C to each point in the bootstrap distribution B, projected on the hyperline connecting C to the new sample spectrum X. In this implementation of $\psi(T,B,X,C)$, constructing a hypothetical plane through C such that the hyperline from X to C is normal to the plane allows these $S_{(p)}$ distances to be given a direction along the hyperline. Points in the bootstrap distribution that are on the same side of the plane as X are assigned positive distances in $S_{(p)}$. The remainder of the elements of $S_{(p)}$ have negative values. This directional assignment can be accomplished by multiplying the elements of $s_{(p)i}$ for which $\{S_{(O2)}^2 + s_{(OR)i}^2 < s_{(2R)i}^2\}$ by -1. At this point the values of $S_{(p)}$ representing points in B that are outside of the hypercylinder are discarded for the remainder of the calculations:

$$S_{(q)} = O(\{s_{(p)i} \mid s_{(R)i} < r_h\}) \qquad (13)$$

and n_h becomes the number of elements in $S_{(q)}$.

For a symmetric 1 SD contour on T, $1 = [0.16n_h]$ and $u = [0.84n_h]$, making the confidence interval along the hyperline connecting X and C $\{s_{(q)1} < C < s_{(q)u}\}$. Note that if n_h is less than about 50, the interval will not be very precise at all. The uncorrected σ can be found by either

$$\frac{(|s_{(q)1}| + |s_{(q)u}|) \, n^{1/2}}{2} \qquad (14)$$

-10-

or by calculating the standard deviation of $S_{(q)}$ and multiplying it by $n^{1/2}$. Once σ is known the distance to the test spectrum in uncorrected BEAST SDs (suitable for unskewed training sets) is simply

$$(\sum_{j=1}^{d} (c_j - x_j)^2)^{1/2} / \sigma \qquad (15)$$

Of course, many training sets are skewed and l and u should be adjusted to compensate for the skew before finding $s_{(q)l}$ and $s_{(q)u}$. At the start of this adjustment one must be aware of the number of replicate points available in B in order to select an adequate contour level for the training set T. For $m \leq 1000$ this contour should probably be one, so $a = \Phi(-1)$. Setting $z_a = \Phi^{-1}(a)$ and locating the center of T by $c_{(T)j} = M_j(t_{ij})$ sets the stage for the adjustment of the confidence limits to compensate for skew.

$C_{(T)}$ will tend to lie in space in the direction opposite to the direction of the skew (with respect to C) because of the leverage effect of skewed points on the mean. This fact is the basis of the confidence-limit adjustment, and the calculation of the magnitude of the adjustment begins with a determination of the distance and direction of the difference between C and $C_{(T)}$ with respect to the hyperline connecting C to X.

$$S_{(COR)} = (\sum_{j=1}^{d} (c_{(T)j} - c_j)^2)^{1/2} \qquad (16)$$

$$S_{(C2R)} = (\sum_{j=1}^{d} (c_{(T)j} - x_j)^2)^{1/2} \qquad (17)$$

-11-

$$S_{(CUB)} = (S_{(O2)} + S_{(COR)} + S_{(C2R)}) / 2 \qquad (18)$$

$$A_c = (S_{(CUB)} (S_{(CUB)} - S_{(O2)}) (S_{(CUB)} - S_{(COR)}) (S_{(CUB)} - S_{(C2R)}))^{1/2} \quad (19)$$

$$S_{(CR)} = 2(A_c) / S_{(O2)} \qquad (20)$$

$$S_{(CP)} = (S_{(COR)}^2 - S_{(CR)}^2)^{1/2} \qquad (21)$$

The directional sign given to $S_{(CP)}$ is opposite that given to $S_{(p)}$. If $\{S_{(O2)}^2 + S_{(COR)}^2 > S_{(C2R)}\}$ then $S_{(CP)}$ is multiplied by -1.

At some point it may be useful to compare the mean of $S_{(q)}$ to the median of $S_{(q)}$. If the two are substantially different $S_{(q)}$ may be skewed. The Central Limit Theorem applies to $S_{(q)}$ so the presence of skew probably indicates that n_h points are not enough to create a stable confidence-limit adjustment. If skew is present two options are available: (1) go back to $\kappa(T)$ and specify a larger m, or (2) increase r_h and recalculate $\psi(T,B,X,C)$ (note that this option may cause a loss of directional selectivity that can bias the quantiles of $S_{(q)}$). Finally, it should be noted that $S_{(q)}$ has been ordered at this point and therefore the use of some common ways of calculating $M(S_{(q)})$ will result in very poor running times for the algorithms. To efficiently find $M(S_{(q)})$, simply select the $(n_h/2+1/2)$th element of $S_{(q)}$ where n_h is odd, and the mean of the $(n_h/2)$th and $(n_h/2+1)$th elements where n_h is even.

In order to make $S_{(CP)}$ perform well as an adjustment in a computational environment where almost any axis scale or skew is possible, $S_{(CP)}$ is replaced by $S_{(CP)}\delta + M(S_{(q)})$. The addition of $M(S_{(q)})$

-12-

helps to assure that the correction $S_{(CP)}$ and and its array-analog z_o (defined below) have the same sign (direction) when $S_{(q)}$ is slightly skewed, and δ provides a skew sensitivity adjustment. Typically δ has a value between 0 and 1 that is set empirically for each combination of T and B to keep the absolute magnitude of the adjustment inside of the actual values of $S_{(q)}$.

The calculation of the z_o adjustment from $S_{(CP)}$ proceeds as follows:

$$f_i = s_{(q)i} - S_{(CP)} \qquad (22)$$

$$I_{(N)} = \{1,\ 2,\ 3,\ \dots,\ n_h\} \qquad (23)$$

$$z_e = [\ R(F(I_{(N)}))\ ] \qquad (24)$$

$$z_o = \phi^{-1}(z_e/\ n_h) \qquad (25)$$

If $|2z_o| > |z_\alpha|$ then δ should be decreased and the calculation resumed at eq. 16. Otherwise, new 1 and u values for $S_{(q)}$ are calculated:

$$1 = [\ \phi(2z_o + z_\alpha)n_h\ ] \qquad (26)$$

$$u = [\ \phi(2z_o - z_\alpha)n_h\ ] \qquad (27)$$

As in the case of the uncorrected BEAST SD σ, the confidence interval along the line connecting X and C is $\{s_{(q)1} < C < s_{(q)u}\}$. In this implementation of the BEAST, the upper confidence limit is always the one

-13-

closest to the test spectrum X.   Thus, σ_c is simply $s_{(q)u}$, and the distance in adjusted BEAST SDs from the training set T to the test spectrum X is

$$(\sum_{j=1}^{d}(c_j - x_j)^2)^{1/2} \, / \, ((\sigma_c/|z_a|)n^{1/2}) \qquad (28)$$

-14-

TECHNICAL REPORT DISTRIBUTION LIST, GEN

	No. Copies		No. Copies
Office of Naval Research Attn: Code 1113 800 N. Quincy Street Arlington, Virginia 22217-5000	2	Dr. David Young Code 334 NORDA NSTL, Mississippi 39529	1
Dr. Bernard Douda Naval Weapons Support Center Code 50C Crane, Indiana 47522-5050	1	Naval Weapons Center Attn: Dr. Ron Atkins Chemistry Division China Lake, California 93555	1
Naval Civil Engineering Laboratory Attn: Dr. R. W. Drisko, Code L52 Port Hueneme, California 93401	1	Scientific Advisor Commandant of the Marine Corps Code RD-1 Washington, D.C. 20380	1
Defense Technical Information Center Building 5, Cameron Station Alexandria, Virginia 22314	12 high quality	U.S. Army Research Office Attn: CRD-AA-IP P.O. Box 12211 Research Triangle Park, NC 27709	1
DTNSRDC Attn: Dr. H. Singerman Applied Chemistry Division Annapolis, Maryland 21401	1	Mr. John Boyle Materials Branch Naval Ship Engineering Center Philadelphia, Pennsylvania 19112	1
Dr. William Tolles Superintendent Chemistry Division, Code 6100 Naval Research Laboratory Washington, D.C. 20375-5000	1	Naval Ocean Systems Center Attn: Dr. S. Yamamoto Marine Sciences Division San Diego, California 91232	1